

Identification and classification of cis-regulatory elements in the amphipod crustacean *Parhyale hawaiiensis*

Dennis A. Sun^{1,*}, Jessen V. Bredeson¹, Heather S. Bruce² and Nipam H. Patel^{2,3,*}

ABSTRACT

Emerging research organisms enable the study of biology that cannot be addressed using classical 'model' organisms. New data resources can accelerate research in such animals. Here, we present new functional genomic resources for the amphipod crustacean *Parhyale hawaiiensis*, facilitating the exploration of gene regulatory evolution using this emerging research organism. We use Omni-ATAC-seq to identify accessible chromatin genome-wide across a broad time course of *Parhyale* embryonic development. This time course encompasses many major morphological events, including segmentation, body regionalization, gut morphogenesis and limb development. In addition, we use short- and long-read RNA-seq to generate an improved *Parhyale* genome annotation, enabling deeper classification of identified regulatory elements. We discover differential accessibility, predict nucleosome positioning, infer transcription factor binding, cluster peaks based on accessibility dynamics, classify biological functions and correlate gene expression with accessibility. Using a Minos transposase reporter system, we demonstrate the potential to identify novel regulatory elements using this approach. This work provides a platform for the identification of novel developmental regulatory elements in *Parhyale*, and offers a framework for performing such experiments in other emerging research organisms.

KEY WORDS: ATAC-seq, *Parhyale hawaiiensis*, RNA-seq, Evolution, Functional genomics

INTRODUCTION

Advances in genomic techniques have facilitated genetic research in emerging research organisms. Genome sequencing has become substantially less expensive over time (<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>), enabling researchers to study genome composition. Moreover, the establishment of genetic perturbation techniques, such as RNAi and CRISPR-Cas9 mutagenesis, has enabled the rapid characterization of gene function in a range of different research organisms. RNAi knockdown has been applied to various insects (Christian et al., 2015; Mito et al., 2011), chelicerates (Sharma et al., 2013), flatworms (Rouhana et al., 2013; Srivastava et al., 2014) and numerous other organisms to study gene function. More recently, CRISPR-Cas9 mutagenesis has enabled both

targeted ablation and targeted transgenesis in organisms as diverse as cephalopods (Crawford et al., 2020) and reptiles (Rasys et al., 2019). Using these tools, it is now possible to identify and perturb the function of genes across diverse organisms to understand how genes and genomes evolve across the tree of life.

Although the study of gene function in diverse systems has grown more tractable, the study of gene regulation has historically proven difficult even in model organisms. Unlike protein-coding genes or noncoding RNAs, which can often be predicted with gene sequence alone, cis-regulatory elements (CREs) are composed of complex assemblies of transcription factor binding sites which can be computationally challenging to identify (Li et al., 2015; Wasserman and Sandelin, 2004). These sites can change dramatically over short periods of time; for example, in the context of the *even skipped* stripe 2 enhancer across several *Drosophila* species, the position of crucial transcription factor binding sites shifted in as short a span as 1-2 million years divergence (Ludwig et al., 1998). Such rapid evolutionary flexibility, although important for driving short- and long-term evolutionary processes, makes the identification of such elements in novel research organisms a considerable challenge.

In recent years, novel techniques for identifying CREs have been developed that rely on the fact that CREs often occur in regions of accessible chromatin. The earliest, including DNase I hypersensitivity and Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq) enabled researchers to locate accessible chromatin regions and discover novel regulatory elements in a range of research organisms (Giresi et al., 2007; Lai et al., 2018; Pérez-Zamorano et al., 2017). However, such techniques were hampered by low signal-to-noise ratios, the requirement for very large amounts of tissue and technically challenging protocols. In 2015, Buenrostro et al. published papers describing the Assay for Transposase-Accessible Chromatin with next-generation sequencing (ATAC-seq), a novel technique using a hyperactive version of the Tn5 transposase, as a way of identifying accessible chromatin regions genome-wide (Buenrostro et al., 2015). In this paper, they demonstrated that ATAC-seq performed comparably with previous accessibility techniques, with reduced tissue requirements and increased convenience and speed.

ATAC-seq has proved to be an effective technique for traditional model systems, and has offered a true revolution for emerging model systems. The low material input requirements of ATAC-seq are of particular use in emerging research organisms, for which generating the millions of cells required for previous techniques would prove challenging. Researchers have applied ATAC-seq to gar embryos in order to study the evolution of limb development (Gehrke et al., 2015) and acoel flatworms and hydra to study regeneration (Cazet et al., 2021; Gehrke et al., 2019), among numerous other examples. Thus, ATAC-seq is now established as a generalizable tool that is particularly well-suited to accelerating work in emerging research organisms.

¹Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720, USA. ²Marine Biological Laboratory, Woods Hole, MA 02543, USA.

³Department of Organismal Biology and Anatomy, University of Chicago, Chicago, IL 60637, USA.

*Authors for correspondence (dennis.a.sun@berkeley.edu; npatel@mbi.edu)

 D.A.S., 0000-0003-1000-7276; J.V.B., 0000-0001-5489-8512; H.S.B., 0000-0003-3199-8224; N.H.P., 0000-0003-4328-654X

In 2017, Corces et al. reported a further-improved version of ATAC-seq, Omni-ATAC-seq, which provides several improvements to the standard ATAC-seq protocol (Corces et al., 2017). In particular, it has been reported to have a greater signal-to-noise ratio, decreased mitochondrial read contamination and applicability to fixed and frozen tissues, with only minor modifications to the buffers and detergents used in the standard ATAC-seq protocol.

In this paper, we apply the Omni-ATAC-seq protocol to embryos of the amphipod crustacean *Parhyale hawaiiensis*, an emerging research organism for the study of arthropod development, evolution and regeneration (Paris et al., 2021 preprint; Stamatakis and Pavlopoulos, 2016; Sun and Patel, 2019). We identify dynamic regions of open chromatin ('peaks') across a broad swath of *Parhyale* developmental time. We comprehensively analyze our Omni-ATAC-seq data to predict the position of nucleosomes along the genome and infer the footprints of transcription factors bound to peaks. Using fuzzy clustering (Kumar and E Futschik, 2007), we partition our peaks into groups based on similar accessibility trajectories, revealing groups of peaks with different transcription factor footprint enrichment and nucleosome occupancy that are active at different points in development. In addition, we use short- and long-read RNA-seq to improve the *Parhyale* genome annotation and investigate the relationship between accessibility and gene expression over time during development.

Parhyale has served as a platform for foundational discoveries about such processes as body plan evolution (Martin et al., 2016), the evolution of arthropod limbs (Bruce and Patel, 2020) and the evolution of regeneration (Konstantinides and Averof, 2014). By facilitating the identification of regulatory elements in this research organism, our work will enable other researchers to investigate the complexities of gene regulatory evolution in these processes and others. Furthermore, by enabling the assessment of cis-regulatory elements in *Parhyale*, we open avenues to investigate fundamental mechanisms of gene regulation in the understudied non-insect crustacean clade.

Although this work primarily focuses on *Parhyale*, our methods can provide examples of how one can perform thorough analyses of ATAC-seq and RNA-seq data using existing tools to generate hypotheses about gene expression dynamics and regulatory element function. Such approaches can be applied to a diverse range of organisms, and will facilitate deeper understanding of gene regulation across the tree of life.

RESULTS

Omni-ATAC-seq identifies open chromatin across *Parhyale* developmental stages

To identify developmental regulatory elements, we performed Omni-ATAC-seq on duplicate libraries of 15 stages of *Parhyale* embryonic development (Fig. 1A-C; Fig. S1A). We evaluated the quality of our libraries using a variety of standard tests, such as fragment size analysis and enrichment of Omni-ATAC signal at promoters (see 'Omni-ATAC sequencing quality control' in Materials and Methods and Fig. 2A,B). Our analyses indicated that Omni-ATAC-seq performed as expected in identifying regulatory elements genome-wide, with low mitochondrial read contamination.

To assess whether our libraries were capable of capturing significant variation over time, we performed principal component analysis (PCA) on our libraries (Fig. 2C). PCA revealed that our libraries were primarily separated along two principal components (PCs), PC1 (60% of variation) and PC2 (18% of variation), with a considerable drop in variation explained in other PCs (Fig. S2D). PC1 appeared to be associated with developmental time, with earlier

developmental stage libraries showing a negative loading and later developmental stages showing a positive loading. Within the PCA plots, samples from the S21 and S22 time points appeared to be more separated in PC2 relative to the other time points. A discussion of this observation and its implications can be found in the supplementary Materials and Methods and Figs S3 and S4C.

We used Genrich (<https://github.com/jsh58/Genrich>) to call significant peaks (merged q -value between replicates < 0.05) in each of our 15 time points independently, and merged overlapping peaks across time points using bedtools (Quinlan and Hall, 2010), yielding 190,078 genomic regions which we used as our 'peaks' in downstream analyses, including the ImpulseDE2 (IDE2) pipeline. The dynamic accessibility of these peaks is illustrated in a heatmap in Fig. 2D. Of these peaks, 163,227 (85.87%) were classified as having statistically significant variation ($\text{padj} < 0.01$) over our time-course; 60,909 (37.32%) were classified as having transient expression dynamics; and 88,231 (54.05%) were classified as showing positively or negatively monotonic expression dynamics (Fig. 2E). These results indicate that we were able to identify many dynamically accessible regions across the *Parhyale* genome.

To identify regions of dynamically accessible chromatin, we used the IDE2 pipeline (Fischer et al., 2018) (see Materials and Methods for a brief summary of the software's advantages). IDE2 produces a fitted model of accessibility for each of the peaks used in the analysis (Fischer et al., 2018) (see Fig. S4C,D for examples). We used these model fits to summarize the global properties of the dynamic peaks we identified. We observed that our model fits tended to achieve maximum accessibility at early, middle and late developmental stages (Fig. 2F). These time points also appeared to be associated with peaks that showed strong loading in PCA (Fig. S2G), indicating that early, middle and late time points are the primary drivers of variation in our dataset.

These global analysis results indicate that our Omni-ATAC-seq experiments captured information normally found in ATAC-seq data, including strong enrichment at promoters. Overall, the low mitochondrial read contamination, large library size and low fraction of duplicated reads suggest that our Omni-ATAC-seq data are of high quality. In addition, the results of the IDE2 differential accessibility analyses indicate that the vast majority of accessible regions in the *Parhyale* genome show dynamic accessibility over developmental time.

Improving the *Parhyale* genome annotation using short- and long-read RNA-seq

Although Omni-ATAC-seq signal showed enrichment across annotated mRNA starts in a genome-wide analysis, careful examination of individual genes and gene models indicated that many of the MAKER gene annotations are fragmented (see Fig. S5A). A more accurate genome annotation would improve both genome-wide analyses and enable the precise classification of candidate regulatory elements as promoters, exonic and intronic regulatory elements, and intergenic regulatory elements. To generate a more complete genome annotation, we performed RNA-seq using two approaches: short-read sequencing using the Illumina NovaSeq platform and long-read sequencing using Oxford Nanopore technology. For four developmental stages (S13, S19, S21 and S23), we generated triplicate RNA-seq libraries (Fig. 3A,B; representative embryo images in Fig. S1B).

We assembled multiple transcriptomes from each of the sequencing approaches (Fig. 3B; see Materials and Methods). To generate a more complete genome annotation, we used the Mikado pipeline to merge our assembled transcriptomes with

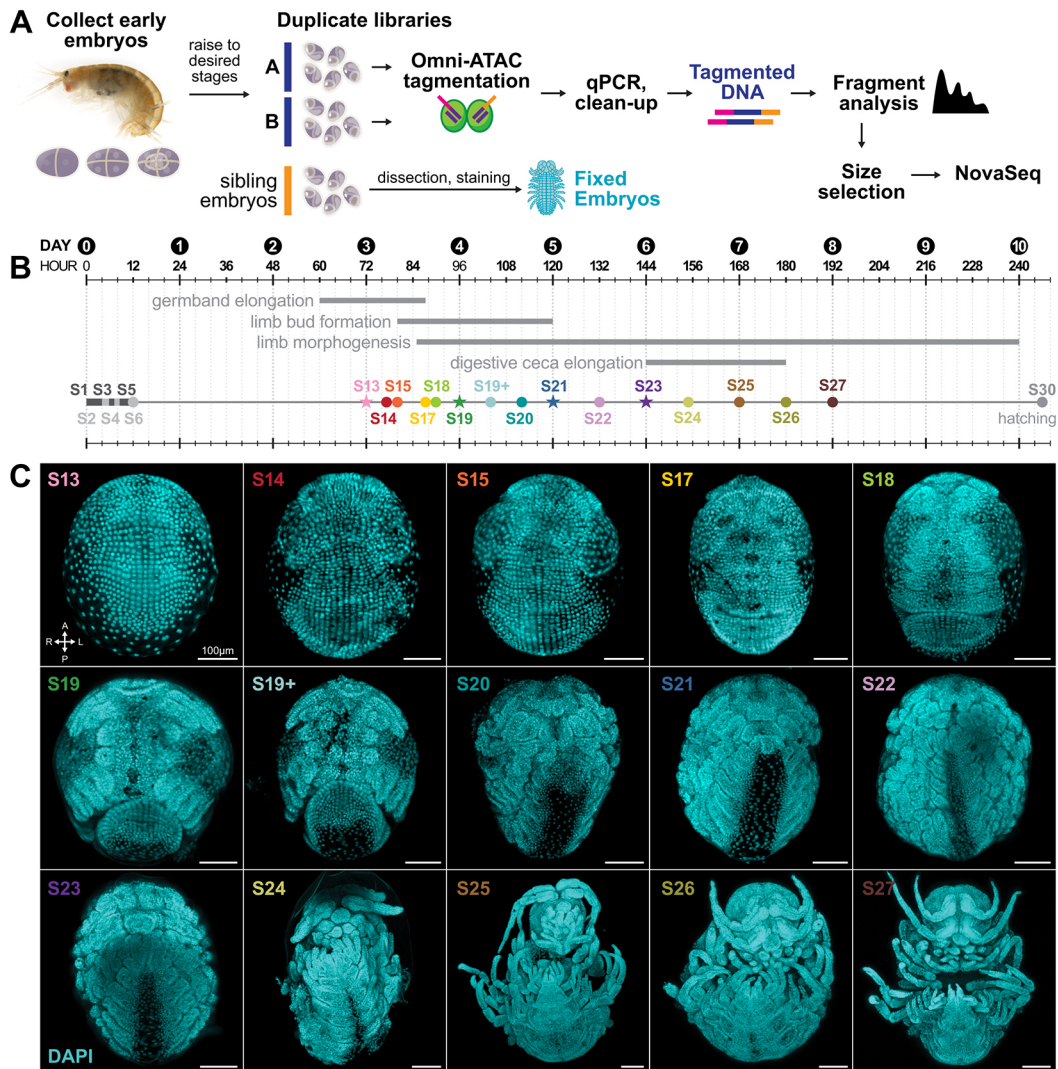


Fig. 1. Time-course ATAC-seq in *P. hawaiensis* embryos. (A) Overview of ATAC-seq protocol. Embryos were collected at early developmental stages and raised to specific developmental time points. Duplicate libraries were generated by tagging five embryos each from a single clutch of sibling embryos. qPCR analysis was used to determine the optimum number of additional cycles of PCR. Tagmented DNA was cleaned and fragment analysis was performed to assess library quality. Final libraries were pooled in equal concentrations, size-selected, and sequence on an Illumina NovaSeq short read sequencer. (B) Timeline of developmental stages. RNA-seq libraries were also generated for time points marked with a star. (C) Representative embryo images from clutches used for Omni-ATAC-seq. Embryos are stained with DAPI and mounted ventral-side up. A, anterior; P, posterior; L, left; R, right. Scale bars: 100 μ m.

the previously-generated Kao et al. (2016) transcriptome and the MAKER genome annotation. The Mikado pipeline and others (e.g. EvidentialGene) use a variety of metrics to compare transcripts from different sources to determine the ‘best’ gene models for each gene region in the genome (Gilbert, 2019 preprint; Venturini et al., 2018). We evaluated the Mikado-merged transcriptome (hereafter ‘Mikado transcriptome’) using BUSCO, and observed a score comparable with the best of the individual transcriptomes (90.9% complete), and a marked improvement compared with the MAKER annotation (80.9% complete) (Fig. 3C).

To further evaluate the completeness of the genome annotation, we examined individual genes and loci of interest. Using previous Rapid Amplification of cDNA Ends (RACE) data, we observed that the Mikado transcriptome has a more complete Hox cluster than the MAKER annotation (Fig. S5A-C), including a more complete *Hox3* gene than has been previously reported. To more comprehensively evaluate gene model completeness between transcriptomes, we generated a custom script to compare transcriptome data

with previous RACE data for additional developmental genes (see Materials and Methods). When we compared the completeness measures between the Mikado transcriptome and the MAKER genome annotation, we observed a marked improvement. Notably, the Mikado transcriptome had a greater proportion of RACE genes with promoter-peak overlap (Mikado 0.84>MAKER 0.71), as well as a greater fraction of ‘single’ transcripts (Mikado 0.73>MAKER 0.55) (Fig. 3D).

To further evaluate the quality of the Mikado transcriptome as a reference, we compared its performance with the MAKER annotation using a variety of metrics (Fig. 3E). The Mikado transcriptome produced a larger number of total gene models ($n=21,218$) than are found in the MAKER annotation ($n=15,105$). Using bedtools, we attempted to assign a candidate promoter peak to each transcript and observed that the Mikado transcriptome had a larger number of transcripts with an assigned promoter (Mikado $n=18,946$; MAKER $n=11,909$), further suggesting improved genome-wide

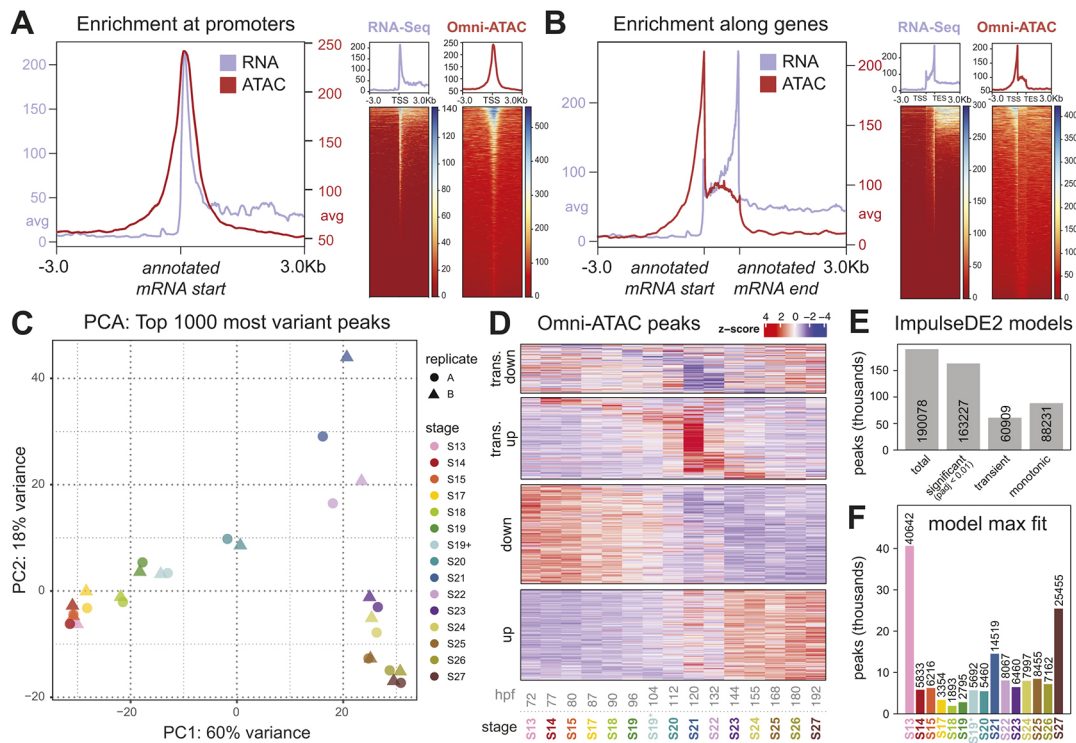


Fig. 2. Whole-genome analysis of ATAC-seq reveals enrichment of promoter signals, dynamic accessibility over developmental time. (A) Omni-ATAC and RNA-seq signal enrichment at promoters. Annotated mRNA start positions are aligned and average signal across genomic positions is plotted for either Omni-ATAC or RNA-seq signal. Axes reflect the average signal of RNA-seq or Omni-ATAC reads across all mRNA start positions in the genome. A strong enrichment of RNA-seq signal is observed 3' of mRNA start sites, whereas enrichment of Omni-ATAC-seq signal appears to be symmetric around mRNA start sites. (B) Omni-ATAC and RNA-seq signal enrichment across gene bodies. Omni-ATAC signal is greatly enriched 5' of gene bodies and slightly enriched across genes, whereas RNA-seq reads are enriched across gene bodies, with an increase in enrichment towards the 3' end of gene bodies. (C) PCA of Omni-ATAC libraries. PCA loadings were assigned to the top 1000 variant peaks, where variance was measured as row variance across all libraries. PC1 (60% of variance) appears to correlate with developmental time. PC2 (18% variance) appears to be associated with middle developmental time points. (D) Heatmap of Omni-ATAC accessibility dynamics, generated from IDE2 model fits. Colors reflect a z-score calculated from IDE2, where red (high z) indicates increase in accessibility and blue (low z) indicates decrease in accessibility. (E) Bar chart indicating number of peaks classified as significant, transient and monotonic by IDE2. (F) Bar chart quantifying stage of maximum accessibility for significant IDE2 model fits. Peaks predominantly achieved maximum accessibility at the start and end of the time course.

gene completeness. We also observed stronger enrichment of Omni-ATAC-Signal at promoters of Mikado genes compared with MAKER genes (Fig. 3F,G), as well as improved performance using genome annotation software (see Materials and Methods). Given these results indicating an improved genome annotation, we used the Mikado transcriptome as our reference for downstream analyses.

Using the Mikado transcriptome as a reference, we assigned Omni-ATAC-seq peaks to a number of different spatial categories ('peak types') along the genome (Fig. 3H; see Materials and Methods). We observed approximately equivalent numbers of peaks over gene bodies (34.4%: 6.2% exonic, 28.2% intronic) and in intergenic regions (37.2%: 7% proximal intergenic, 30.2% distal intergenic). We further partitioned the intergenic peaks into proximal and distal segments, with distal intergenic peaks representing those peaks >10 kb away from the nearest gene (see Materials and Methods for rationale for this cutoff). The majority of intergenic peaks in our dataset were distal intergenic peaks (82%; 30.2% distal intergenic / 37.2% total intergenic peaks), indicating that many intergenic regulatory elements could not have been identified using previous approaches. The average distance of intergenic peaks was 73,351 bp away from the nearest gene (Fig. S4B). Nearly a fifth of regulatory elements (18.2%) were also located on contigs that did not contain genes, and were classified as 'unknown'.

As our peaks represent the merged combination of peaks across individual time points, we also evaluated the proportion of peak

types from each time point (Fig. 3H). Peaks analyzed in this way could reveal differences in the spatial distribution of peaks across the genome over development. After normalizing for the number of 'unknown' peaks, the proportion of peaks belonging to intergenic and promoter regions declined slightly as development progressed, whereas the proportion of gene body peaks increased (Fig. S4E). There appeared to be slightly more intergenic peaks than gene body peaks at all developmental time points, and the ratio of intergenic peaks to gene body peaks declined very slightly over time (Fig. S4F). These data indicate that intergenic and gene body peaks have different enrichment trajectories over time. As development progresses, gene body peaks are increasingly enriched, and may have a greater impact on gene regulation.

Inferring nucleosome positioning and transcription factor footprints from Omni-ATAC-seq

ATAC-seq identifies regions of open chromatin. However, with sufficient depth of sequencing, one can infer additional information from the pattern of insertions of ATAC-seq adapters into the genome. Using our Omni-ATAC data, we predicted the temporal positioning of nucleosomes near Omni-ATAC peaks using NucleoATAC and inferred the footprints of transcription factors bound to accessible peaks using HINT-ATAC.

NucleoATAC enables the prediction of nucleosome positioning from ATAC-seq data (Schep et al., 2015). We applied NucleoATAC

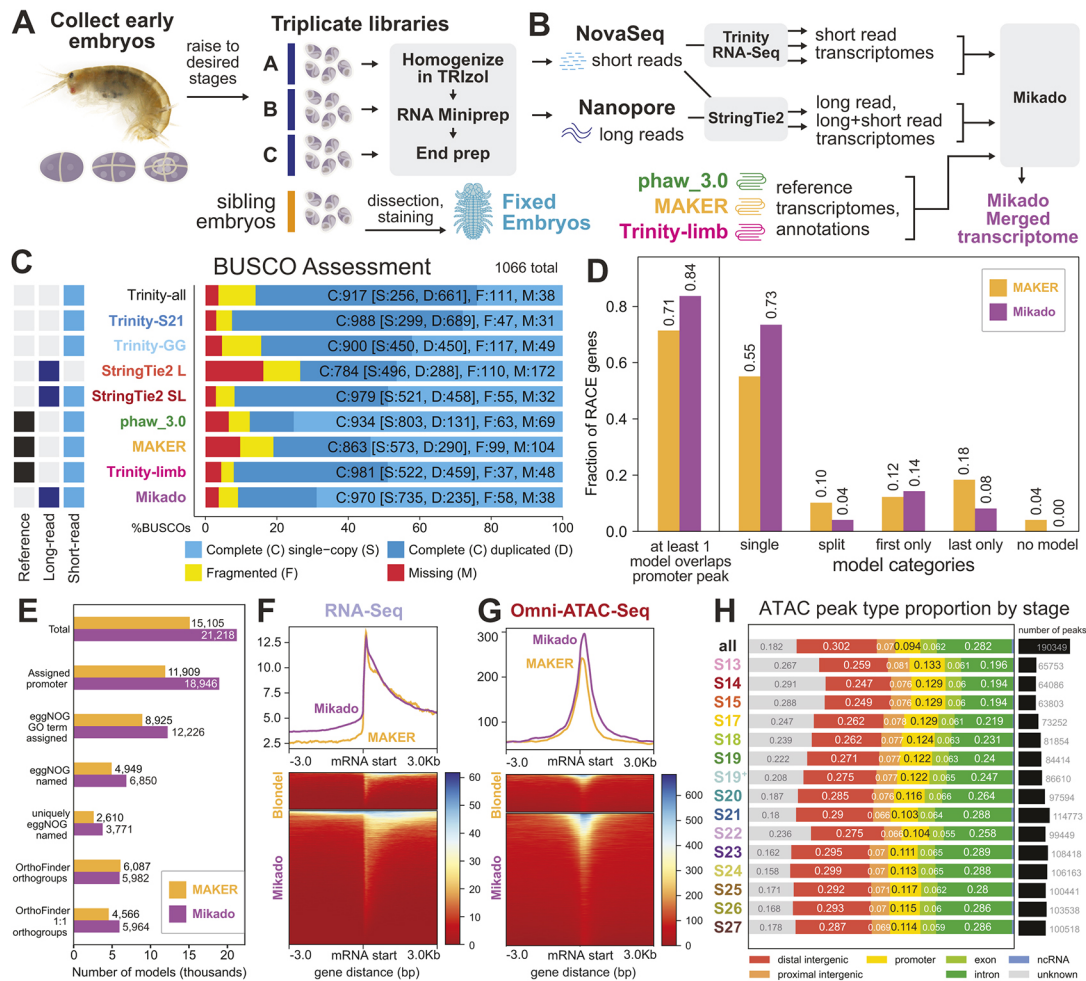


Fig. 3. Time-course RNA-seq of *Parhyale* embryos. (A) RNA-seq protocol overview. Triplet libraries were generated for four developmental stages. (B) RNA-seq transcriptome assembly and merging pipeline. Short-read, long-read and long+short-read transcriptomes were generated for pooled developmental stages matched to Omni-ATAC time points (S13, S19, S21, S23) using Trinity and StringTie2. Transcriptomes were also assembled using Trinity for short reads from a limb developmental time course (see Materials and Methods). Assembled transcriptomes were then merged with additional transcripts from previous publications into a Mikado transcriptome. (C) BUSCO scores for representative datasets and Mikado transcriptome. Overall, the StringTie2 short+long-read transcriptome (StringTie2 SL) performed comparably with the best Trinity transcriptomes. The Mikado transcriptome appeared to have a comparable BUSCO complement to the other datasets with high BUSCO completeness. (D) Gene model completeness for Mikado transcriptome versus MAKER genome annotation based on a dataset of 49 RACE genes. The Mikado transcriptome appears to have more complete gene models and less fragmentation. (E) Comparison of number of gene models of different functional annotation categories between MAKER genome annotation and Mikado transcriptome. The Mikado transcriptome appears to produce more annotated genes with a higher annotation quality. (F) Plot comparing RNA-seq signal pileups between MAKER mRNA starts and Mikado mRNA starts. Axis represents mean signal at mRNA start sites for each dataset. (G) Plot comparing Omni-ATAC-seq signal pileups between MAKER mRNA starts and Mikado mRNA starts. (H) Categorization of Omni-ATAC-seq peaks in each stage-specific ATAC-seq library by their position relative to Mikado gene models. Distal intergenic peaks are defined as those >10 kb away from the nearest gene, whereas unknown peaks were peaks that were located on a config that did not have a gene model.

to a window of ± 500 bp around our Omni-ATAC-seq peaks at each developmental stage. NucleoATAC depends on a stereotypical signal of Tn5 insertion around nucleosomes referred to as a 'V-plot', which resembles the fragmentation pattern around chromatin subjected to chemical fragmentation. We observed across developmental time points a clear V-plot signal generated by NucleoATAC (example of V-plot from S21 libraries shown in Fig. S8A).

A well-established signal found across eukaryotes is the presence of strongly positioned +1 and -1 nucleosomes around gene promoters, and nucleosome depletion at RNA-Pol II binding sites (referred to as nucleosome-free region; NFR) (Brogaard et al., 2012; Radman-Livaja and Rando, 2010). We visualized NucleoATAC signal and nucleosome occupancy around mRNA starts genome-wide, and observed strong signals associated with +1 and -1

nucleosomes (Fig. 4A,B; Fig. S8B-D). We also observed a depletion of nucleosomes and a decrease in NucleoATAC signal between the +1 and -1 nucleosomes, which may represent NFRs. This signal was observed consistently in each of our stage-specific libraries (Fig. S8E,F). These data indicate that we were able to identify the stereotypical nucleosome signal at promoters, suggesting that NucleoATAC is able to infer nucleosome positions within our dataset genome-wide.

Deeply sequenced ATAC-seq data also enables transcription factor binding inference. We used HINT-ATAC to identify transcription factor footprints in Omni-ATAC-seq peaks from each developmental stage (Li et al., 2019) and evaluated the enrichment of these transcription factor footprints across developmental time. We used the JASPAR database of position frequency matrices (PFMs) to comprehensively identify possible

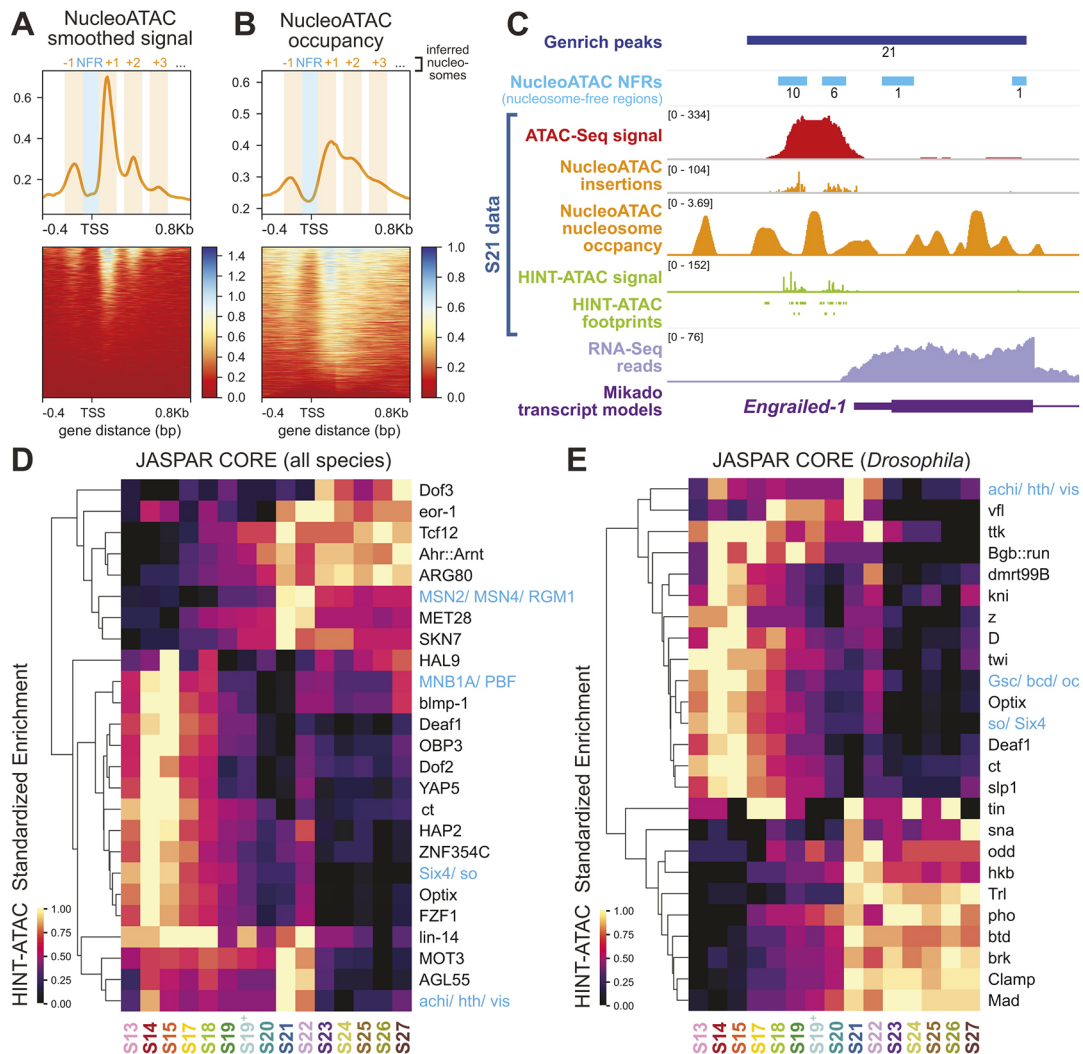


Fig. 4. Inference of nucleosome positioning and transcription factor binding from Omni-ATAC-seq data. (A) NucleoATAC smoothed signal at mRNA starts genome-wide. A clear pattern of peaks in the signal can be observed, suggestive of nucleosome positioning. Axis reflects mean signal across all mRNA starts at stage S21. (B) Inferred nucleosome occupancy at mRNA starts genome-wide. Axis reflects mean signal across all mRNA starts at stage S21. Inferred nucleosome annotations on A and B added based on observed signal. (C) Visualization of NucleoATAC, HINT-ATAC, Omni-ATAC and RNA-seq data at the inferred *Engrailed-1* promoter at stage S21. Deeper inference on Omni-ATAC-seq data allows insight into the broader chromatin landscape at developmental time points included in this dataset. (D) Summary of standardized enrichment of JASPARD CORE position frequency matrices (PFMs) among HINT-ATAC identified footprints across developmental stages. The top 25 most-enriched PFMs are displayed. Color is based on standardized enrichment, where minimum and maximum enrichment within each row are set to 0 and 1, respectively. PFMs marked in blue had identical enrichment ratios within each group at all time points, likely due to having highly similar PFMs. Overall, PFMs appeared to be enriched at early, late or middle developmental stages. (E) Summary of top 25 most-enriched *Drosophila* PFMs from the JASPARD CORE database. Overall, PFMs appeared enriched at early, late or middle developmental stages.

binding motifs (Fornes et al., 2020). We observed that HINT-ATAC-enriched transcription factor footprints appeared to form three groups based on the timing of highest enrichment: early, S21/S22 and late (Fig. 4D,E). These results are consistent with the general trends observed in PCA, which suggest that the early, middle and late time points in development have the greatest variation.

The data generated from these analyses enable the visualization of chromatin accessibility, predicted nucleosome position, inferred transcription factor binding and RNA-seq expression at individual loci across around half of *Parhyale* development. The strengths of these data are illustrated in Fig. 4C at the predicted promoter of the *Parhyale Engrailed-1* locus for stage S21. Thus, for any genomic region of interest, one can develop a comprehensive prediction of the local chromatin environment at each of the 15 time points in our dataset.

Identifying gene regulatory programs from Omni-ATAC-seq using fuzzy clustering

To assess the potential for our dataset to provide new biological insights, we attempted to identify distinct gene regulatory programs based on differential accessibility. We used the Mfuzz package to perform fuzzy c-means clustering on the matrix of read counts generated from our Omni-ATAC-seq peaks (Kumar and E Futschik, 2007). We determined that our peaks could be optimally partitioned into nine clusters (see Materials and Methods).

We evaluated the differences between clusters through a variety of approaches (Figs 5 and 6). Fig. 5D shows the standardized raw accessibility scores of peaks in each of the nine clusters over time. Our clusters appeared to show substantial differences in the timing of their maximum accessibility, the GO terms associated with genes located near peaks in each cluster and also the enrichment of binding footprints for different transcription factor families. For example,

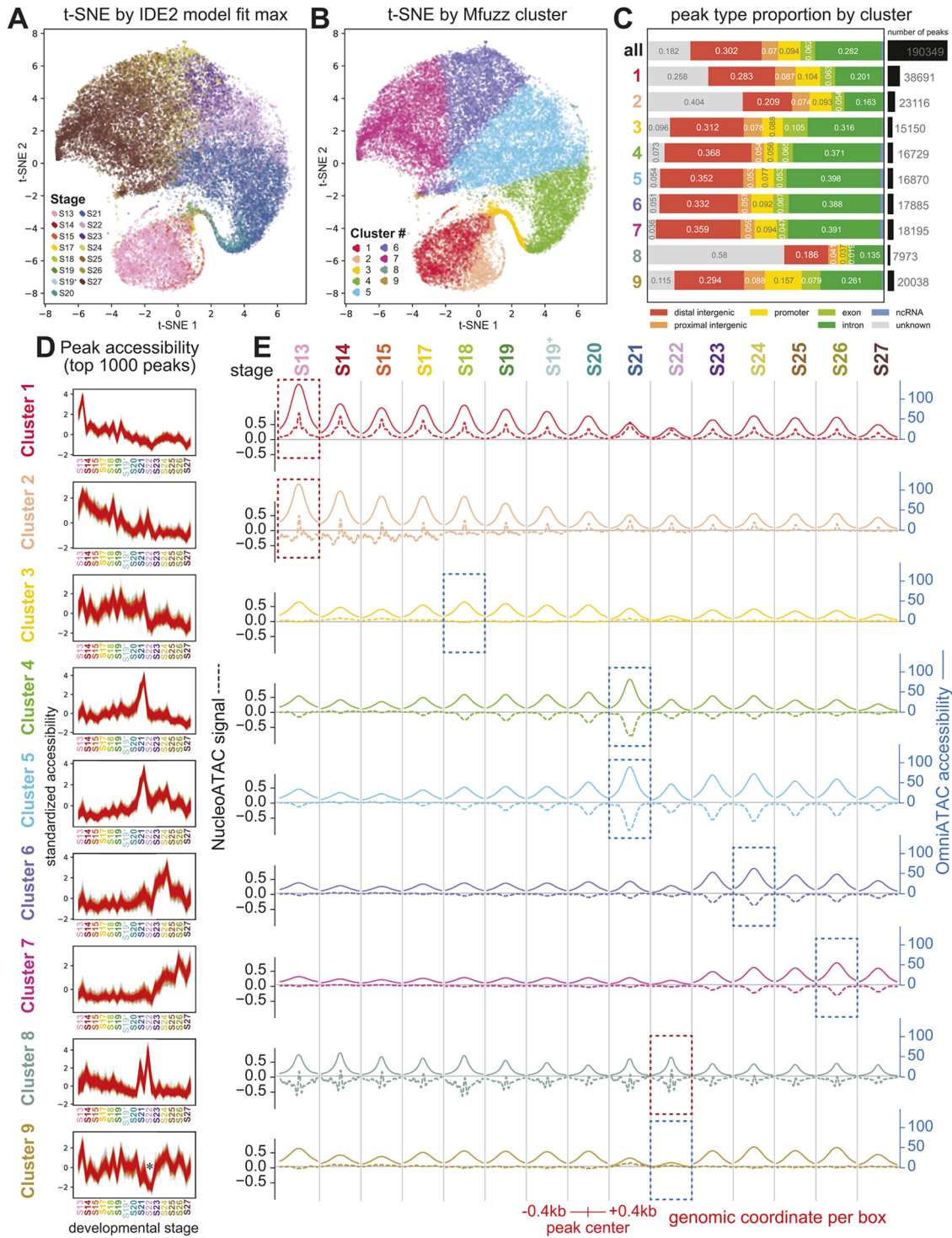


Fig. 5. Identification and classification of regulatory element clusters. (A) t-SNE plot of the top 10,000 most statistically significant peaks identified by IDE2, colored by IDE2 max fit. (B) t-SNE plot of the same points from A colored by Mfuzz cluster. (C) Distribution of peak position categories by Mfuzz cluster. Clusters 1, 2 and 8 appeared to be enriched for ‘unknown’ peaks, whereas other clusters appeared to be enriched for distal intergenic peaks. (D) Standardized accessibility plot for top 1000 peaks with strongest membership in each of the nine clusters. Each library is plotted as a separate point along the line plot. Line color indicates the cluster membership value for each peak in each plot, with dark red indicating strong cluster membership and blue indicating weaker cluster membership. An asterisk in the Cluster 9 panel marks the decrease in accessibility observed commonly in Cluster 9 peaks, more clearly shown in the IDE2 model fits in Fig. S9B. (E) Line plots showing Omni-ATAC accessibility (solid line, blue axis) and NucleoATAC signal (dashed line, black axis) for peaks in each cluster across time. Each box contains a line plot summarizing each signal at all peaks within that cluster with respect to a given developmental Omni-ATAC-seq library. The center of each line plot reflects the average signal at the center of all Omni-ATAC-seq peaks in that cluster at that developmental stage. Signal is visualized at 0.4 kb upstream and downstream of peak centers. Within each signal type, the axes across line plots for each cluster are identical. Dotted lines around subsections of the plot highlight regions of interest. Dotted blue lines indicate the time point during which peaks in a given cluster achieved both maximum accessibility and decreased nucleosome occupancy. Dotted red lines indicate the time point during which peaks in a given cluster achieved both maximum accessibility and increased nucleosome occupancy.

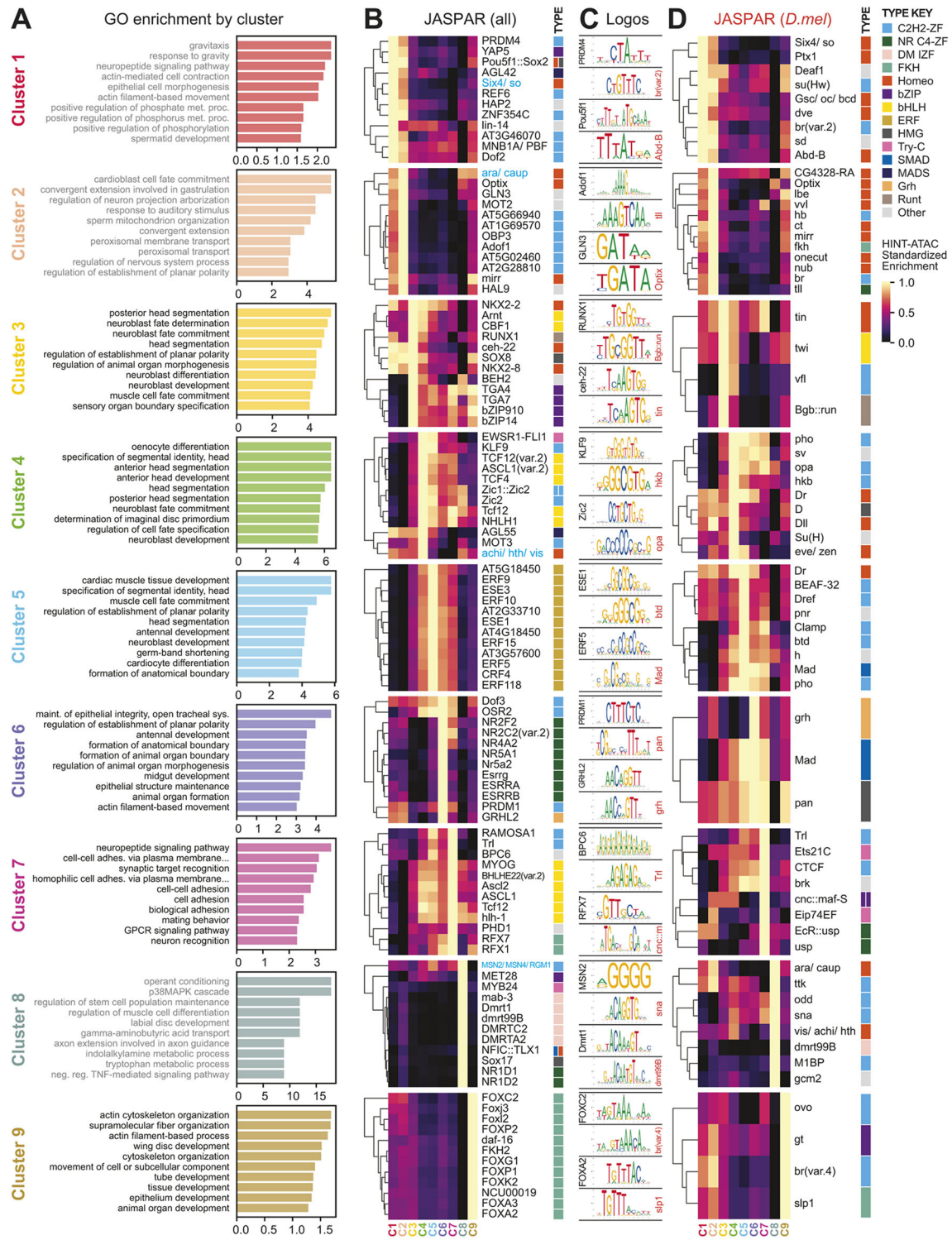


Fig. 6. Functional annotation of clusters using GO enrichment and transcription factor binding prediction. (A) GO enrichment per cluster based on the nearest gene for each peak. For each peak in each cluster, we identified the nearest gene and extracted the *Drosophila* OrthoFinder gene name for that gene. We performed GO enrichment using *Drosophila* GO terms, using the list of all OrthoFinder gene names found in our genome as the background for the analysis. Bar charts display fold enrichment for each GO term. Gray text represents GO terms that did not have a significant *P*-value ($P < 0.05$) after false discovery rate (FDR) correction. (B) HINT-ATAC enrichment per cluster for JASPAR CORE PFMs. For each cluster, we identified all unique transcription factor footprints across all 15 time points found within peaks of that cluster and performed enrichment relative to a randomly-selected background. We show the top 12 PFMs with the greatest enrichment for each cluster that also had the highest enrichment in that cluster relative to other clusters. Colored boxes represent the transcription factor family of each transcription factor in the clustermap. (C) JASPAR motif logos for select pairs of transcription factors between non-*Drosophila* and *Drosophila* transcription factor with similar motifs as determined by STAMP alignment. For each cluster, two pairs of transcription factors are shown: a non-*Drosophila* transcription factor (labeled in black) and a *Drosophila* transcription factor (labeled in red). In some cases, a strong similarity is observed between the non-*Drosophila* and *Drosophila* PWMs. (D) HINT-ATAC enrichment per cluster for *Drosophila* JASPAR CORE PFMs. Some clusters had fewer than 12 PFMs for which that cluster had the highest enrichment.

Cluster 9 appeared to be associated with peaks that decreased in accessibility in the middle of development (Fig. 5D, asterisk), and genes located near peaks in this cluster appeared to show GO enrichment for cytoskeletal terms. For a fuller description of the differences between clusters, see supplementary Materials and Methods.

We also evaluated transcription factor footprint enrichment for each cluster. Given that many JASPAR PFMs come from non-arthropod sources, we used STAMP to align non-*Drosophila* to *Drosophila* motif sequences based on similarity (Mahony and Benos, 2007) (see Fig. S11). Selected pairs of motifs are shown in Fig. 6C for each cluster (non-*Drosophila* labeled in black, *Drosophila* labeled in red). For some pairs, we were able to identify candidate *Drosophila* transcription factors that appeared highly similar to non-*Drosophila* sequences. For example, in Cluster 9, we observed strong enrichment for FKH family transcription factors, the PFMs of which matched the br(var.4) and slp1 PFMs. This suggests that peaks found in Cluster 9 may be regulated by br(var.4) and slp1 binding in *Parhyale*.

Altogether, our results indicate that clustering Omni-ATAC data using accessibility can identify groups of peaks with similar accessibility trajectories, which can be further analyzed using nucleosome occupancy, GO enrichment and transcription factor footprint enrichment to understand possible biological functions and genetic mechanisms behind differential accessibility.

Concordant and discordant expression and accessibility dynamics appear across development

To understand the relationship between peak accessibility and gene expression, we examined the relationship between the change in accessibility of individual peaks and the change in expression of nearby genes (Fig. 7). We used DESeq2 to evaluate the log₂-fold change of both expression (RNA-seq) and accessibility (ATAC-seq) for all genes and peaks. Among our differential accessibility and expression analyses, we observed both ‘concordant’ and ‘discordant’ relationships. Peak-gene pairs with concordant accessibility and gene expression were those in which the sign of the log₂-fold change in expression and accessibility were in agreement – for example, peaks for which an increase in accessibility was observed concurrently with an increase in expression. Meanwhile, peak-gene pairs with discordant accessibility and gene expression were those for which an increase in accessibility was observed concurrently with a decrease in expression, or vice versa.

We observed both classes of peak-gene pairs among all peaks, but also among promoter peaks. Over time, the number of concordant peak-gene pairs increased from 1888 at the S19 versus S13 comparison, to 5738 at the S21 versus S19 comparison, and then decreased to 5240 at the S23 versus S21 comparison (Fig. S12D). Meanwhile, the number of discordant peaks gradually increased over time (761 to 2231 to 3115 peak-gene pairs) (Fig. S12D). These trends were also observed for the promoter-only peaks. The gradual increase in discordant peaks may indicate an increase in repressive gene regulation as gene expression becomes refined over the course of differentiation.

Overall, the classification of peaks as concordant or discordant in accessibility and gene expression may provide downstream users of this data with hypotheses about CRE function. Given that CREs have recently been shown to function as both activating elements and silencers, depending on tissue context (Gisselbrecht et al., 2020; Halfon, 2020), we cannot directly map concordant peaks to enhancers or discordant peaks to silencers. Without information about higher-order chromosome contacts, it is also difficult to

precisely assign a given regulatory element’s function to a particular gene of interest.

However, information about concordance and discordance of peaks could be useful for researchers deciding which among many peaks surrounding a gene of interest could be most fruitful for reporter construction. To facilitate analysis of accessibility and expression for peak-gene pairs of interest, we have included functions enabling visualization of these two factors for arbitrary peak-gene pairs, illustrated in Fig. S13. The visualization tool displays accessibility and gene expression for user-selected Omni-ATAC peaks and Mikado genes over RNA-seq stages (S13, S19, S21, S23), fold change in accessibility and expression between adjacent time points, significance of fold change as evaluated by DESeq2 and concordance/discordance assignment at each time point.

Omni-ATAC-seq re-identifies known *Parhyale* regulatory elements

To assess the practical usefulness of our Omni-ATAC-seq dataset, we compared our Omni-ATAC-seq peaks to known regulatory elements in *Parhyale*. A very limited set of regulatory elements have been described in *Parhyale*: a muscle reporter, PhMS (Pavlopoulos and Averof, 2005); a heat shock element, HS2a (Pavlopoulos et al., 2009); an embryonic ubiquitous reporter, PEB (Liubicich, 2007); and two reporter constructs for *Parhyale Opsin-1* and *Opsin-2* (Ramos et al., 2019). We were unable to locate any Omni-ATAC-seq peaks for the *Parhyale Opsin* genes, likely owing to their late expression (outside of our developmental time course) and very low cell number (a handful of cells per eye). However, strong Omni-ATAC-seq enrichment was observed at the PhMS, HS2a and PEB elements, as illustrated in Fig. S14.

Each of these reporter constructs consists of the first exon and intron of a gene. In each case, we observed an Omni-ATAC peak overlapping the 5’ end of the gene. For the PhMS reporter (Fig. S14B), previous work had identified a cluster of putative basic helix-loop-helix (bHLH) transcription factor binding sites. These sites were captured within our Omni-ATAC peak. The HS2a reporter has been reported to contain two binding sites for heat shock factor (HSF) proteins upstream of a minimal promoter (PhHsp70), both of which are captured within our Omni-ATAC peak (Fig. S14D). Finally, the PEB element appears to contain two Omni-ATAC peaks. These results suggest Omni-ATAC is able to identify the position of functional CREs, and has been able to capture those regions important for CRE function.

Minos transposase reporter assays can reveal novel promoters and distal enhancers

To assess the function of candidate regulatory elements identified by Omni-ATAC-seq, we employed a *Minos* transposase reporter assay (Fig. 8A). In this assay, we injected *Minos* transposase mRNA along with a transposon donor plasmid containing a reporter gene construct that expressed DsRed to one- and two-cell embryos. Once per day, from 3 to 10 days postfertilization (dpf), we screened for DsRed expression on a Zeiss LSM780 confocal microscope. We tested a variety of different reporters, summarized in Table 1 (see Materials and Methods for a discussion of candidate reporter selection approaches and see supplementary Materials and Methods and Fig. S16 and Fig. S17 for description of spontaneous expression patterns observed).

Among the novel reporter constructs we tested, two showed robust expression: pMi(Hsc70-4) and pMi(ne1)-Hsp70-p2. The pMi(Hsc70-4) construct was constructed using the putative

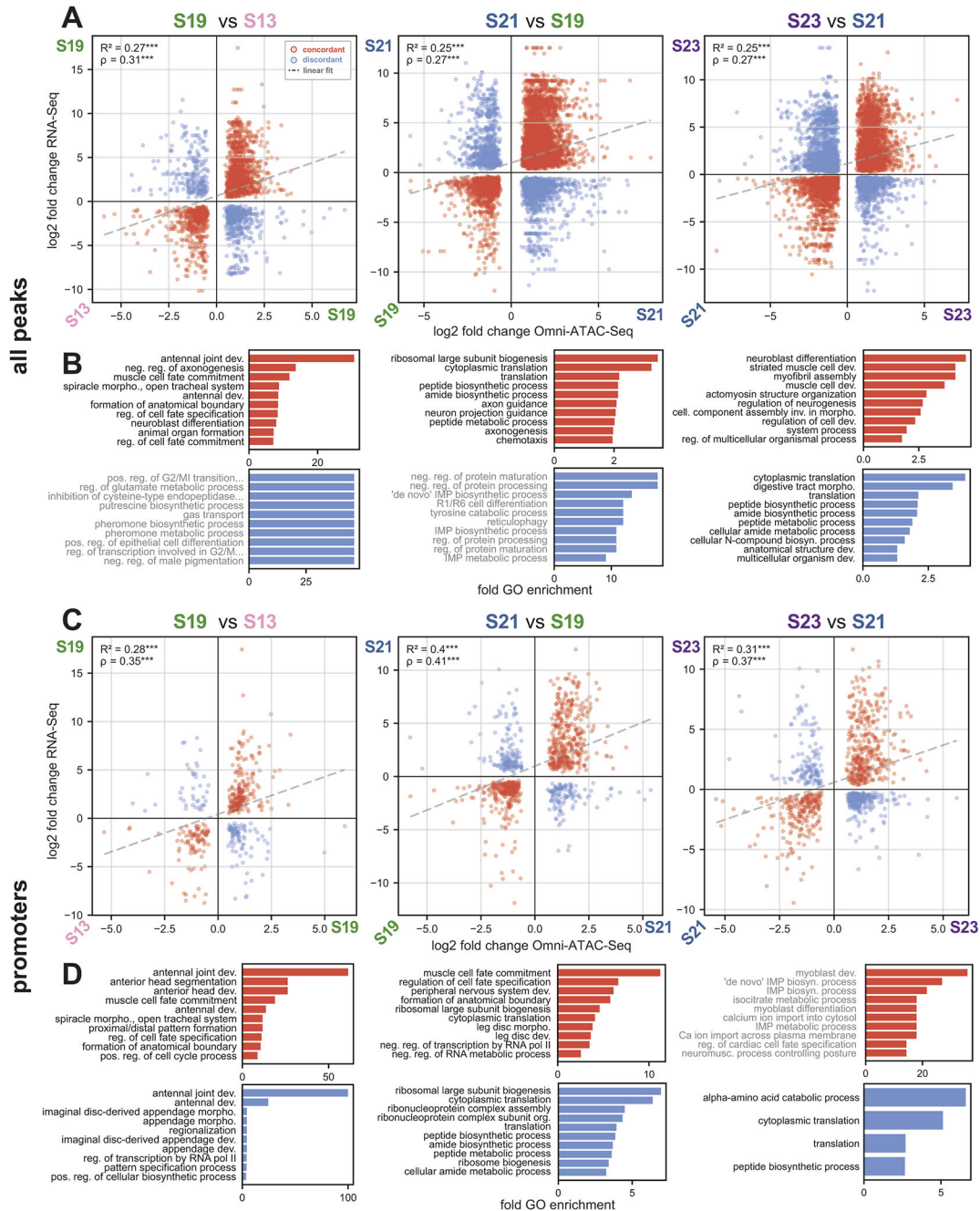


Fig. 7. Correlation between RNA-seq and Omni-ATAC-seq. (A) Correlation between \log_2 -fold change RNA-seq and \log_2 -fold change Omni-ATAC-seq at peak-gene pairs in sequential pairwise comparisons of developmental stages. Only peak-gene pairs with a significant difference in expression and accessibility between the two time points are plotted. Positive axis reflects higher expression or accessibility at the later developmental stage for each plot. Points colored in red reflect concordant relationships between \log_2 -fold change in RNA-seq and Omni-ATAC-seq; blue peaks reflect discordant relationships. Dotted line represents a linear fit to all data in the plot. Pearson correlation R^2 and Spearman correlation ρ for all points are displayed in each plot. $^{***}P < 0.001$. (B) GO-term enrichment for gene-peak pairs with concordant or discordant expression and accessibility \log_2 -fold change for all peaks. Gene lists were extracted based on the nearest Mikado gene to each peak. (C) Correlation between \log_2 -fold change RNA-seq and \log_2 -fold change Omni-ATAC-seq at all peaks in each developmental stage. Positive axis reflects higher expression at the later developmental stage for each plot. Pearson correlation R^2 and Spearman correlation ρ for all points are displayed in each plot. $^{***}P < 0.001$. (D) GO-term enrichment for gene-peak pairs with concordant or discordant expression and accessibility \log_2 -fold change, for promoter peaks.

promoter peak of the Hsc70-4 gene (Fig. 8B,D). This reporter showed strong expression in a cluster of neurons associated with the eye beginning at ~ 9 dpf (Fig. 8C). The expression pattern of these neurons appears to be distinct from those observed using a synthetic 3XP3 enhancer, which has previously been used as a positive marker of transformation *Parhyale* (Pavlopoulos and Averof, 2005).

Thus, it is possible to build new reporters from our Omni-ATAC peaks using putative promoter peaks.

The second reporter that showed robust expression was the pMi(ne1)-Hsp70-p2 construct, which contains the PhHsp70 minimal promoter along with a strong peak extracted from the Hsp70 cluster, about 30 kb away from the location of the minimal

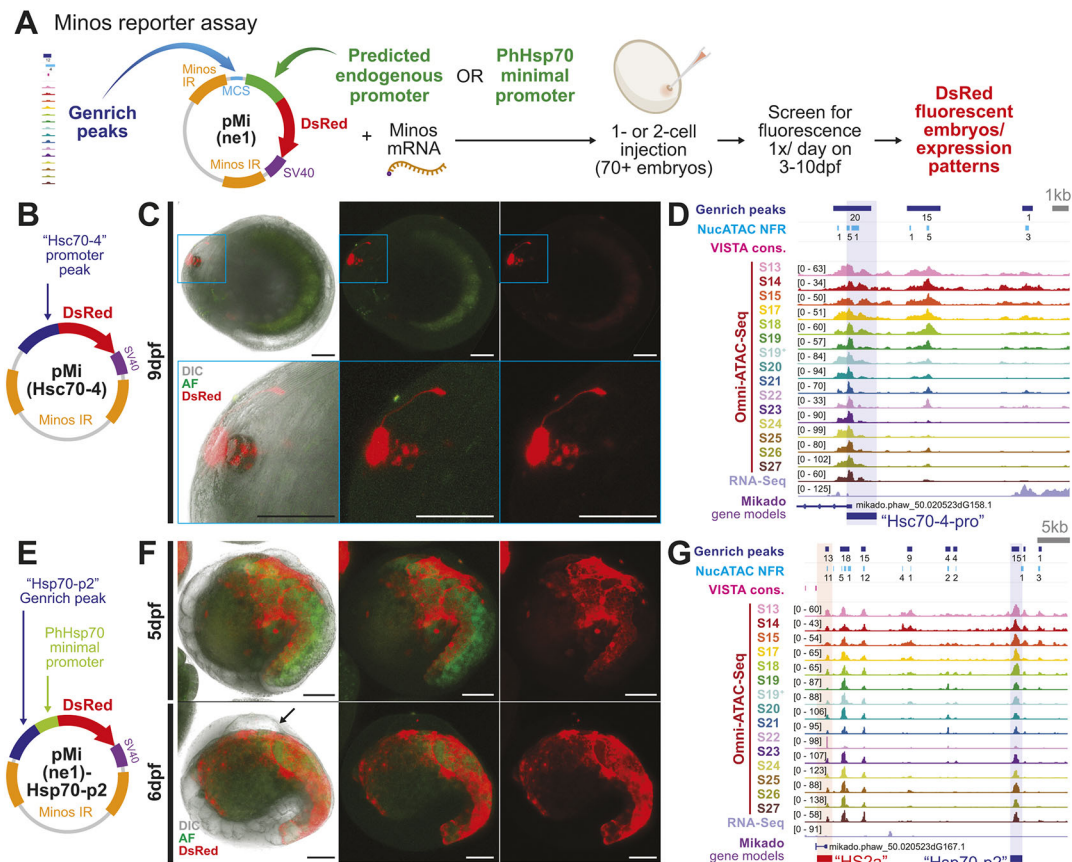


Fig. 8. Minos transposase reporter assay identification of novel regulatory elements. (A) Minos transposase reporter assay and screening approach. (B) Plasmid schematic for the pMi(Hsc70-4) reporter. (C) Expression of the pMi(Hsc70-4) reporter in neurons associated with the eye of a 9 days postfertilization embryo. Bottom row shows a magnified image of the head region. A clear projection of the neurons can be seen. (D) The genomic region of the Hsc70-4 promoter peak. (E) Plasmid schematic for the pMi(ne1)-Hsp70-p2 reporter. (F) Expression of pMi(ne1)-Hsp70-p2 reporter at 5 and 6 dpf showing strong expression in the yolk. Arrow marks aberrant morphology at the dorsal side of the embryo observed consistently in all embryos with fluorescent expression. (G) The genomic region of the Hsp70-p2 peak relative to the HS2a element. The PhHsp70 minimal promoter used in this construct is located within the HS2a element. Scale bars: 100 μ m.

promoter (Fig. 8E,G). Hsp70-p2 drives strong expression in cells of the embryonic yolk (Fig. 8F). Some yolk cells labeled by this reporter moved dynamically around the yolk, whereas others appeared static and showed spongiform morphology (Movie 1). We observed defects in the development of the dorsal portion of the embryo in mid-stage (~6-7 dpf) embryos injected with this construct, which stereotypically displayed a separation between the embryo proper and the eggshell (Fig. 8F, bottom panel, marked with arrow). Although we were unable to establish a genetic line using this construct, we observed strong and reproducible expression in many embryos. This reporter is the first indication that distal enhancers exist and can be identified in *Parhyale*.

Although it remains challenging to identify novel CREs in *Parhyale*, our results suggest that functional regulatory elements can be identified from our Omni-ATAC data. Among the numerous peaks identified in our dataset, we expect that many will reveal novel gene regulatory dynamics. Future work to identify such elements will require the optimization of reporter expression and transgenic strategies to increase the throughput and sensitivity of reporter gene assays.

DISCUSSION

New genomic data enable deeper understanding of underlying biology. This is especially true in emerging research organisms, for which data and resources are limited. This work provides a

wealth of genomic resources for the amphipod crustacean *P. hawaiiensis*. In addition to providing Omni-ATAC-seq and RNA-seq data for a broad developmental time course, our work generates multiple new transcriptomes, an updated genome annotation, a catalog of dynamically accessible chromatin regions and predictions of nucleosome occupancy and transcription factor binding. Moreover, our thorough analysis of regulatory element dynamics through clustering, GO enrichment, predicted transcription factor binding and correlations with RNA-seq present numerous hypotheses for CRE function. Such data will support researchers in the growing *Parhyale* community in efforts to identify and characterize developmental regulatory elements, and provide the foundations for more advanced approaches, including single-cell sequencing techniques, which rely on high-quality reference datasets.

The *Parhyale* genome contains many distant and dynamic regulatory elements

From our data, we were able to glean new information about the global dynamics of gene regulation during *Parhyale* development, as well as the composition of the *Parhyale* genome. We observed that the vast majority of our peaks showed dynamic accessibility over developmental time, and that peaks could have a variety of different temporal dynamics, including transient increases and decreases in accessibility alongside more absolute increases and decreases. These results indicate that much of the *Parhyale* genome

Table 1. Plasmids tested using Minos transgenesis

Plasmid	Type	Gene – <i>Drosophila</i> expression	Number injected	Number survived	Number of DsRed or GFP+	Phenotype
pMi(gm)	Promoter	N/A – <i>Parhyale</i> muscle reporter PhMS	109	39	24	Green muscle expression
pMi(retn)	Promoter	<i>retained</i> – expressed in some neurons	240	104	5	No consistent phenotype
pMi(SVP)	Promoter	<i>seven up</i> – expressed in ventral neuroblasts	122	31	0	None
pMi(su-s-2)	Pro+peak	<i>suppressor of sable</i> – embryonic ubiquitous expression	127	57	7	No consistent phenotype
pMi(Rx)	Pro+peak	<i>retinal homeobox</i> – expressed in various neural tissues	74	28	0	None
pMi(Hsc70-4)	Promoter	<i>Heat shock 70 complex protein 4</i> – embryonic ubiquitous	212	99	22	Neurons associated with the eye
pMi(En1)	Promoter	<i>engrailed</i> – parasegment stripes	71	33	0	None
pMi(En1)-p1	Pro+peak	"	67	46	0	None
pMi(En1)-p1.5	Pro+peak	"	70	51	0	None
pMi(En1)-p5	Pro+peak	"	68	58	1	Faint expression domain
pMi(Sp69)	Promoter	<i>Sp69</i> – leg gap gene	74	56	0	None
pMi(Sp69)-p1	Pro+peak	"	50	30	0	None
pMi(Sp69)-p1.5	Pro+peak	"	86	32	3	No consistent phenotype
pMi(Sp69)-p2.5	Pro+peak	"	71	35	0	None
pMi(Sp69)-p3	Pro+peak	"	241	121	8	Head, leg, gill
pMi(ne1)-Hsp70-p1	Pro+peak	Hsp70 complex	45	26	0	None
pMi(ne1)-Hsp70-p2	Pro+peak	Hsp70 complex	310	81	61	Yolk cells

N/A, not applicable; ", same as previous row.

undergoes dynamic changes in accessibility over developmental time.

With an improved genome annotation, we were also able to determine that many (~37.2%) regulatory elements in *Parhyale* are located in intergenic regions, and that among all elements at all time points, elements located >10 kb away from the nearest gene (distal intergenic peaks) were the largest group, followed by intronic peaks. These results are notable, given that many of the currently studied protostome genomes are small (for example, *Drosophila melanogaster* is 180 Mb, *Tribolium castaneum* is 200 Mb and *Bombyx mori* is 530 Mb). In such organisms, regulatory elements tend to occur within short distances from promoters, and it is common in such systems to attempt to build reporter genes using a short window upstream of the promoter of a gene of interest.

Parhyale is an example of an arthropod with a large genome (3.6 Gb, or ~10% larger than *Homo sapiens* at 3.2 Gb). Given the large genome size, and the large proportion of elements located distant to genes, we expect that regulatory elements within *Parhyale* may be located more distantly from gene promoters than in other arthropods. Future attempts at building reporter genes in this organism will need to account for the large number of distant regulatory elements, which may be crucial to proper reporter expression.

Our data are among the first to identify regulatory elements genome-wide in a non-insect arthropod (Gatzmann et al., 2018; Kissane et al., 2021), and also examine one of the largest sequenced arthropod genomes available to date (Kao et al., 2016). These data will enable other researchers to examine the relationship between genome size and regulatory element composition across more diverse taxa, and will be useful in developing a deeper understanding of how genome organization affects gene expression across the metazoans.

Combined short- and long-read transcriptomes can improve genome annotations

Our work can also serve as a guide for other researchers working in emerging research organisms. First, our work demonstrates the utility of combining short- and long-read sequencing in generating more accurate genome annotations for emerging research

organisms. Even with a small number of sequences (~1.2 million reads) generated using inexpensive Nanopore sequencing, we were able to assemble a transcriptome with a moderate BUSCO score (71%). Combining short reads and long reads (StringTie2 SL), we were able to construct a transcriptome with both a high BUSCO score (91.8%) and low rate of gene fragmentation.

Moreover, the completeness of the StringTie2 SL transcriptome met or exceeded that of several of our Trinity-derived transcriptomes. Examination of individual genomic regions in the *Parhyale* genome indicates that Trinity generates numerous spurious transcripts, which can be confounding for whole-genome analysis (see Fig. S6 for examples). Thus, for other researchers with access to a genome assembly, we would strongly recommend performing both short- and long-read sequencing, and assembling using multiple different assembly strategies to generate higher quality genome annotations. For those without access to a high-quality genome, we would caution against relying strictly on transcripts generated from software such as Trinity, which, in our system, generated many transcripts that did not appear to match with previous RACE data. Filtering transcripts by expression values, as well as employing transcriptome-merging pipelines such as Mikado (in the case of a high-quality genome) and EvidentialGene (in the absence of a genome), would likely help remove such spurious transcripts.

Deep analysis of ATAC-seq data can reveal more than just dynamic accessibility

Although the most direct analysis of ATAC-seq data can enable the identification of dynamic regulatory elements, our work uses additional tools to infer nucleosome positioning and transcription factor binding from our ATAC-seq data. Using NucleoATAC, we were able to recover clear signals of nucleosome positioning at promoters, a quality observed in numerous organisms. Using HINT-ATAC, we were able to predict transcription factor binding across developmental stages. Such analyses enable a glimpse of the possible chromatin landscape of the genome, particularly in the absence of more direct and specific, but also more time-consuming and expensive, approaches such as ChIP-seq or CUT&RUN.

In addition to performing inference, we were able to identify clusters of peaks with potentially distinct biological functions using fuzzy clustering. These clusters of peaks appeared to differ in their accessibility dynamics, nucleosome positioning and enrichment of transcription factor footprints, indicating that meaningful biological differences can be gleaned from deeper analysis of ATAC-seq data. For example, we observed strong enrichment in Cluster 9 for FKH transcription factor footprints matching the PFMs for *Drosophila* br(var.4) and slp1. The strongest peaks in this cluster appeared to share a common decrease in accessibility during stages S21-S22, and an enrichment for genes with GO terms related to cytoskeletal function. This group of peaks appeared to show a decrease in accessibility at a time point during which the embryo undergoes a dramatic morphogenetic event in which it splits along the midline (Browne et al., 2005). We hypothesize that FKH domain-containing transcription factors may play a role in that important morphogenetic event. Using our analyses as a starting point, other *Parhyale* researchers may be able to make hypotheses about their own developmental processes of interest.

Omni-ATAC-seq identifies old and new regulatory elements, including enhancers

Our data were able to capture most of the previously-identified *Parhyale* regulatory elements. In addition, we were able to demonstrate that our data contain new regulatory elements. Notably, we demonstrate the first identification of a distal regulatory element, Hsp70-p2, located about 30 kb from the PhHsp70 minimal promoter. These results are the first identification of an enhancer separated by a large distance from a minimal promoter element in this organism.

Our results demonstrate that Omni-ATAC-seq is able to identify novel regulatory elements, and we expect that numerous new reporter genes will be built from these data. Previous work relied on examining candidate regulatory elements within individual genomic regions, or attempting to build new reporters through random integration events. With this dataset, it is now possible to identify candidate regulatory elements for any gene of interest and to screen for expression patterns using the Minos reporter system. Moreover, by evaluating concordance and discordance of gene expression and accessibility using this dataset, it is possible to further refine the list of candidates for reporter construction based on the potential of individual peaks to drive or repress gene expression.

We were unable to identify clear reporters from some of the developmental genes we tested, such as *Sp69*; however, many peaks near these genes remain to be tested. Future work to build new reporters in *Parhyale* may require the development of more efficient transgenesis and screening strategies. For example, improving the efficiency of CRISPR-mediated homologous recombination or CRISPR-mediated non-homologous end-joining transgenesis could enable researchers to insert a reporter near a candidate regulatory region of interest. Moreover, future approaches will also need to account for the numerous distant regulatory elements in the *Parhyale* genome, which may be important for gene regulation. Examining local DNA interactions using approaches such as Hi-C or other chromatin conformation capture approaches will be instrumental in identifying distant regulatory regions.

Together, these approaches have enabled the exploration of chromatin dynamics and transcription factor binding in an emerging model organism. Our work illustrates how a single, easily adapted protocol can yield data amenable to deep analysis. We would recommend that other researchers using ATAC-seq or similar assays in emerging model organisms also take advantage of the additional

information that could be gleaned from deep analysis of their data, which could provide further insights into how both local and global changes to nucleosome occupancy and transcription factor binding influence their biological processes of interest.

Conclusion

By combining Omni-ATAC-seq with RNA-seq across a broad developmental time course, our work is able to identify and classify numerous candidate CREs in the genome of the amphipod crustacean *P. hawaiensis*. We demonstrate how deep analysis of Omni-ATAC data can facilitate the identification of peaks with distinct accessibility, nucleosome occupancy and transcription factor footprint enrichment. We further classify peaks as concordant or discordant regulatory elements by integrating differential accessibility and differential expression, revealing potential relationships between regulatory elements and nearby genes. Moreover, we show the potential to identify novel reporter genes using candidate promoters and enhancers from our data.

This work provides a substantial resource to the *Parhyale* community, and should accelerate the study of gene regulation in this emerging research organism. In addition, our work can serve as a framework for other researchers deploying ATAC-seq and RNA-seq approaches in emerging research organisms. Through deep analysis of ATAC-seq data, combined short- and long-read sequencing, and integration of accessibility and gene expression, such researchers can identify regulatory elements with distinct biological functions and advance the study of gene regulation in their organisms of interest.

MATERIALS AND METHODS

Crustacean cultures

Parhyale hawaiensis of the Chicago-F isolate were raised at 25°C and fed a diet of carrots, shrimp pellets and Spirulina flakes in Tropic Marin artificial seawater with a salinity of 31-35 ppm in plastic tanks.

Embryo cultures

Embryos were collected using previously described protocols (Rehm et al., 2009), staged using the Browne et al. (2005) staging guide, and raised at 27°C in a humidity-controlled incubator in filter-sterilized artificial seawater. For ATAC-seq and RNA-seq experiments, clutches of 15 or more embryos were collected and staged between S1 and S6, as these time points are among the shortest and most morphologically identifiable during early development.

Embryo staging

Two groups of five embryos each from a single clutch were used in the ATAC-seq experiments, and three groups of five embryos each from a single clutch were used in the RNA-seq experiments. Embryos of the correct stage were selected based on morphological characteristics as described in the Browne et al. (2005) staging guide, and any embryos with abnormal or asynchronous morphology were discarded. Morphologically representative embryos from the same clutch as used for the ATAC-seq were photographed on ventral and lateral positions within 10 min of beginning the tagmentation procedure. Any remaining embryos were boiled briefly and fixed using the protocol described in Rehm et al. (2009) and DAPI stained to further confirm staging. The process of staging embryos, imaging embryos, boiling and mashing embryos for Omni-ATAC tagmentation were all performed within a 30 min time interval for each developmental stage.

RNA-seq

Embryos for RNA isolation were homogenized in TriZol using a DWK Life Sciences (Kimble) Biomasher II Closed System Disposable Tissue Homogenizer, and RNA was isolated using the Zymo Direct-Zol Miniprep Plus kit. RNA quality was assessed based on fragment analysis using an Agilent 2100 Bioanalyzer. cDNA was generated from RNA using

the TaKaRa SMART-seq v4 Ultra Low-Input kit. cDNA was then sequenced using the Illumina NovaSeq to generate short reads. Using the Nanopore Direct cDNA barcoding kit (SQK-LSK109), cDNA was also sequenced on a Nanopore MinION flow cell to generate long reads.

Limb development RNA-seq methods

A pool of embryos consisting of embryonic stages 19, 20, 22, 23, 25 and 28 were homogenized with DWK Life Sciences Kimble Kontes Pellet Pestle Cordless Motor in Trizol and extracted using Trizol. PolyA+ libraries were prepared with the Truseq V1 kit (Illumina), starting with 0.6–3.5 mg of total mRNA, and sequenced on the Illumina HiSeq 2000 as paired-end 100 base reads, at the QB3 Vincent J. Coates Genomics Sequencing Laboratory.

Limb development RNA-seq *de novo* transcriptome assembly

For the Trinity-limb (old) assembly, transcripts were assembled *de novo* using Trinity r2013_08_14 (Grabherr et al., 2011) with parameters `-JM 170G -CPU 10 -inchworm_cpu 6 -min_kmer_cov 2 -min_contig_length 49 -group_pairs_distance 700`. Input RNA-seq reads were treated as paired-ends. Output transcript assemblies shorter than 200 bp were discarded. The remaining assemblies were screened for contaminants with BLASTX (BLAST+ v2.2.26; parameters: `-num_descriptions 50 -num_alignments 50 -evaluate 1e-5 -lcase_masking -soft_masking true -seg yes`) (Camacho et al., 2009) against a database of all bacterial proteins downloaded from NCBI (retrieved 2013-08-31), then against Swiss-Prot UniProt human sequences (retrieved 2013-08-31) (The UniProt Consortium, 2019). All hits were further required to cover a minimum of 40% of each assembled transcript sequence. When filtering for human contaminants, a 98% identity threshold was also required. For the Trinity-limb (new) assembly, transcripts were assembled using Trinity v2.5.1 using standard settings.

Transcriptome assemblies

Using the Trinity (Haas et al., 2013) pipeline to assemble the short-read sequences, we generated four developmental stage-specific *de novo* transcriptomes (Trinity-S13 to Trinity-S23) and two merged transcriptomes, one *de novo* (Trinity-all) and one genome-guided (Trinity-GG). In addition, we generated two transcriptomes with reads from several developmental stages covering the time course of limb development [Trinity-limb (new) and Trinity-limb (old)]. Although the merged transcriptomes were generated from all four developmental stages (Trinity-all, Trinity-GG), they scored lower compared with stage-specific transcriptomes when coding DNA sequence completeness was assessed using BUSCO (Fig. 3D). One possible explanation for this could be an increase in transcript assembly fragmentation from de Bruijn graph assembly caused by the high heterozygosity previously described of the *Parhyale* genome (Kao et al., 2016). For long-read sequencing, we mapped the reads to the phaw_5.0 genome using minimap2 and assembled a transcriptome using StringTie2 (Kovaka et al., 2019; Li, 2018). We also used StringTie2 to generate a combined transcriptome containing both short and long reads. Although the long-read transcriptome (StringTie2 L) yielded a low BUSCO score, the short+long-read transcriptome (StringTie2 SL) scored comparably with our other transcriptomes (Fig. 3D).

Short-read RNA-seq data was used to generate both *de novo* and genome-guided transcriptomes using Trinity. For genome-guided assembly, reads were mapped using HISAT2. Long-read RNA-seq data was mapped to the most recent *Parhyale* genome (phaw_5.0) using minimap2 and assembled using StringTie2. A combined transcriptome using both HISAT2-mapped short reads and minimap2-mapped long reads was also generated using StringTie2. See supplementary Materials and Methods ('Trinity transcriptome assembly parameters' and 'StringTie2 transcriptome assembly parameters') for additional information about assembly parameters.

Mikado transcriptome

Short-read Trinity transcriptomes and the transcriptome from Kao et al. (2016) were mapped to the phaw_5.0 genome using GMAP. Short-read transcriptomes and long-read StringTie2 transcriptomes were merged using the Mikado software along with a previous genome annotation generated by

Leo Blondel (Harvard University, MA, USA) (using MAKER). See supplementary Materials and Methods for additional information about Mikado parameters.

Gene model completeness analysis

To quantify the level of gene model fragmentation in our dataset, we generated a series of manual gene annotations based on RACE sequences. Among 143 previously-generated RACE transcripts, we selected 49 multi-exonic transcripts that appeared to have a single promoter based on the Omni-ATAC-seq data (Fig. S5D). For each of these RACE transcripts, we manually annotated the extent of the first and last exon by comparing RNA-seq read pileups to RACE data and the current genome annotation. We also identified the Omni-ATAC-seq peak most likely to capture the gene promoter, based on the strength of the peak (as evaluated by the number of time points over which we observed a statistically significant peak), as well as overlap to RNA-seq read data. We used these manual annotations (see Table S2) to evaluate the gene models from each of the different transcript sources (transcriptomes or gene annotations).

For each of the 49 genes in our dataset, we evaluated whether any models in each of the transcript sources in our dataset overlapped the promoter peak (Fig. S5E). The ability to unambiguously identify promoters is essential for downstream analyses, including building reporter constructs or targeting CRISPR guides to the 5'-most end of genes. Among the transcript sources, the Trinity *de novo* transcriptomes had the highest proportion (0.96) of genes for which at least one model overlapped with the promoter peak. However, the Trinity models we observed often (0.98, 48/49 RACE genes examined for the Trinity-limb transcriptome) contained numerous spurious transcript fragments in introns, exons and 3' untranslated region (UTR) (see Fig. S6 for summary statistics and examples). Among the remaining transcript sources, the Mikado transcriptome had the highest fraction of genes for which at least one model overlapped with the promoter peak (0.84).

In addition, we assessed the degree of fragmentation of gene models across the 49 RACE transcripts (Fig. S5D). For each gene, we determined first whether any single gene model spanned the first and last exon, representing a complete 'single' transcript. If there was not a single model, we next assessed whether two separate models overlapped with the first and last exon, or if all models only overlapped either the first or last exon. These results together formed a measure of transcript completeness. Comparing the different transcript sources, the StringTie2 SL transcriptome had the highest fraction (0.76) of complete 'single' transcripts, with the Mikado transcriptome having a slightly lower fraction of 'single' transcripts (0.73) (Fig. S5E).

Functional annotation of transcripts

To assess the quality of the two genome annotations, we performed automated functional annotations to assign gene names and functions. We used two approaches: eggNOG and OrthoFinder (Emms and Kelly, 2019; Huerta-Cepas et al., 2019). eggNOG is a rapid and lightweight genome annotation software that assigns gene names, KEGG pathway information and GO terms, among numerous other metrics, to gene models. OrthoFinder facilitates the identification of orthogroups between provided peptide libraries, enabling automated comparisons between gene lists in species of interest.

We observed that the Mikado transcriptome included a greater number of GO-term assigned genes, eggNOG named genes and eggNOG uniquely-named genes than were found in the MAKER annotation (Fig. 3E). We used OrthoFinder to assign orthogroups between the Mikado or MAKER transcriptomes and the list of all *D. melanogaster* peptides from the UNIPROT-SWISSPROT database. The Mikado and MAKER transcriptomes produced similar numbers of orthogroups; however, the Mikado transcriptome produced a greater number of 1:1 orthogroups, which proved useful for downstream GO term enrichment analysis (summarized in Fig. 3E; see Fig. S7 for further explanation of orthogroup size comparisons). See supplementary Materials and Methods ('eggNOG annotation' and 'OrthoFinder annotation') for additional information about functional annotation.

ATAC-seq data

We performed conventional ATAC-seq as per Buenrostro et al. (2015) for stages S13-S22. Data were used for benchmarking in comparison to Omni-ATAC data, but were not used for downstream analyses.

Omni-ATAC-seq tagmentation

Tagmentation was performed using the reagents described in Corces et al. (2017) Omni-ATAC-seq paper, with the following modifications. Instead of the Illumina Nextera TD buffer, 2× Tagmentation Buffer from Wang et al. (2013) was used. Homemade Tn5 enzymes were purified and received as a gift from Jase Gehring (University of California, Berkeley, CA, USA; California Institute of Technology, CA, USA) and assembled into Tn5 transposomes as per Picelli et al. (2014), using adapters purchased from IDT with 5' phosphorylation (Picelli et al., 2014). Before adding RSB+D, embryos were washed 3× using 1× PBS. Embryos were mashed into a near-uniform solution using the tip of a low-retention p10 pipette, and the pipette was visually inspected for any remaining debris. We used the Qiagen MinElute kit for purification and concentration of DNA.

Omni-ATAC-seq library preparation

Saturation PCR conditions for Omni-ATAC-seq libraries were performed using a Roche Lightcycler 480 as per Corces et al. (2017). Optimal conditions for additional amplification after pre-amplification were determined based on the number of additional PCR cycles corresponding to one-third of the maximum value, rounded up. Barcodes were added to samples using primers supplied by QB3-Berkeley using PCR as described in Corces et al. (2017). Following amplification, adapters were removed from libraries using Ampure bead purification and analyzed using a Bioanalyzer or Fragment Analyzer machine to assess library quality. Final libraries were pooled together to have relatively equal proportions based on additional qPCR quantification and size-selected using a Pippin Prep to remove fragments greater than 1.5 kb in size. All libraries were sequenced on one lane of each of Illumina NovaSeq SP 150PE and NovaSeq S1 150PE. Adapters used for libraries are listed in Table S1.

Omni-ATAC-seq quality control

We employed a battery of standard tests used to assess the quality of ATAC-seq data (Fig. S2A). Omni-ATAC-seq libraries were also compared with previous ATAC-seq libraries, and had greater library size, higher numbers of non-duplicated reads and lower amounts of mitochondrial DNA contamination (Fig. S2B,C). Fragment size analysis of Omni-ATAC-seq reads after mapping revealed distinct 1-nucleosome and 2-nucleosome peaks, indicating that tagmentation was efficient (Fig. S2E,F).

To determine whether our Omni-ATAC-seq data represented genuine enrichment in accessible chromatin regions, we examined promoter accessibility genome-wide. A hallmark of successful ATAC-seq experiments is strong enrichment of reads in gene promoters (Yan et al., 2020). We evaluated our ATAC-seq using the most recent phaw_5.0 gene annotations (obtained from Leo Blondel, hereafter referred to as the 'MAKER annotation') at the start of annotated mRNAs and across mRNA lengths, and compared our results with RNA-seq read pileups (Fig. 2A,B). As expected, Omni-ATAC-seq signal is enriched symmetrically at mRNA starts, whereas RNA-seq reads are enriched 3' of mRNA starts. In addition, Omni-ATAC-seq data show greater enrichment at promoters than over annotated mRNA regions, and decreased enrichment outside of annotated mRNAs. These data suggest that our Omni-ATAC-seq performed as expected in identifying promoter regions genome-wide.

Adapter trimming was performed using Trim Galore, which leverages Cutadapt and FASTQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>; https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/; Martin, 2011). FASTQC was also performed before and after adapter trimming to confirm removal of sequencing and Tn5 adapters. Percentage of reads remaining after deduplication was estimated based on FASTQC metrics. Library size was estimated by multiplying raw read count by percentage of reads remaining after deduplication for each lane, and then

summing the two lanes. Reads were aligned to the phaw_5.0 genome as well as the *Parhyale* chrM using Bowtie2, and percentage read mapping was determined using Bowtie2 output (Langmead and Salzberg, 2012). After peak calling, merged peaks were used to evaluate correlation between replicates. Replicates were highly correlated, with a mean Pearson correlation of 0.988 and a mean Spearman correlation of 0.936 (Fig. S2H,I).

Omni-ATAC peak calling

Aligned reads were quality-filtered with a *q*-value of 10 in samtools and used for Genrich analysis. Genrich analysis was run on both duplicate libraries simultaneously; Genrich performs peak calling on each peak individually, and then merges the *P*-values of the replicates using Fisher's method to generate a *q*-value, obviating the need to calculate an irreproducible discovery rate. Output .bedgraph-like files from Genrich were reformatted using a custom Python script to be standard .bedgraph files, which were converted to bigWig files using the bedGraphToBigWig executable from the University of California, Santa Cruz Genome Tools software bundle. Output .narrowPeak files from Genrich were converted into .bed files for ease of visualization in Integrative Genomics Viewer using a custom Python script (https://github.com/mezarque/Parhyale_Genome_Resources).

IDE2 concept and analysis

IDE2 differs from other software for differential expression analysis in that it allows the investigation of trajectories of dynamic expression over large numbers of time points. It does so by modeling a gene expression trajectory as an 'impulse' function that is the product of two sigmoid functions (Chechik and Koller, 2009; Yosef and Regev, 2011). This approach enables the modeling of a trajectory of gene expression in three parts: an initial value, a peak value and a steady state value, thus summarizing an expression trajectory using a fixed number of parameters. With the ability to capture the differences between early, middle and late expression values for each gene in a dataset, IDE2 also enables the detection of transient changes in gene expression or accessibility during a time course. Identifying differential expression over large numbers of time points is difficult for more categorical differential expression software such as edgeR and DESeq2, which generally use pairwise comparisons between time points to assess change over time (Love et al., 2014; Robinson et al., 2010).

Assigning spatial categories to Omni-ATAC peaks

We first assigned the nearest peak within 5 kb of the first 200 bp of each gene (mRNA and ncRNA) as a promoter peak using bedtools closest. We then assigned the remaining peaks into categories based on their position relative to mRNA and ncRNA annotations. Peaks that overlapped with mRNA and ncRNA annotations were assigned as exonic or intronic regulatory elements. The remaining peaks – those which had not been classified as promoters, and which did not overlap with genes – were classified as intergenic peaks. The intergenic peaks were divided into two categories: proximal and distal intergenic peaks. Proximal peaks were those <10 kb away from the nearest gene, whereas distal intergenic peaks were those >10 kb away from the nearest gene. We established this cutoff with the rationale that peaks beyond this distance would be considerably more difficult to isolate as single fragments combined with a promoter peak using PCR, agnostic of their orientation with respect to the promoter element. Such peaks could not have been easily identified using previous approaches, and thus differ from 'proximal' peaks by their necessary identification using genomic techniques.

Fuzzy clustering concept and analysis using Mfuzz

Fuzzy clustering differs from categorical clustering approaches, such as hierarchical clustering and k-means clustering, in that it assigns a probability for each element in the dataset to fall into each of the identified clusters. For example, rather than a given peak falling in either cluster A or B exclusively, a given peak might have a 95% chance of falling in cluster A and a 5% chance of falling in cluster B.

The Mfuzz package allows the specification of two parameters: *c*, the number of clusters to create from the data, and *m*, the stringency of the

clustering. To determine the optimal number of clusters in our fuzzy c-means clustering, we varied the c-value from 3 to 13, and evaluated the quality of clustering using the cluster overlap metric provided by Mfuzz, as well as the silhouette score metric (Fig. S10) (Kumar and E Futschik, 2007). We determined that nine clusters yielded the highest mean silhouette score, while avoiding clusters below the silhouette score average and minimizing size variability between clusters. Fig. 5A plots with t-stochastic neighbor embedding (t-SNE; implemented in scikit-learn) (Hinton and Roweis, 2003; Van der Maaten and Hinton, 2008; Pedregosa et al., 2011) the top 10,000 most significant peaks called by IDE2, where each point is colored by its IDE2 model max fit, while Fig. 5B illustrates the same clusters, colored by their Mfuzz cluster. Fig. 5C indicates the spatial categories of peaks in each cluster.

Differential expression and accessibility analyses

We performed differential expression analysis using IDE2 and DESeq2 using standard settings. To generate the Omni-ATAC-seq read count matrix, we used bedtools multicov, using the merged Genrich peaks as our regions of interest. For our IDE2 analyses using the Omni-ATAC-seq data, IDE2 model fits were extracted from the IDE2 output and used to visualize model expression using a custom Python script (see supplementary Materials and Methods; ‘DESeq2 parameters’ and ‘ImpulseDE2 analysis’). To generate the RNA-seq read count matrix for DESeq2, we generated a gene_trans_map file for the Mikado transcriptome, as would be available for the Trinity RNA-seq analysis pipeline, and used the built-in Trinity differential expression pipeline (align_and_estimate_abundance.pl, abundance_estimates_to_matrix.pl) with Kallisto to generate a matrix of read counts. To comply with the requirement for integer counts in DESeq2 analysis, we rounded each value to the nearest whole number.

NucleoATAC nucleosome predictions

Quality-filtered reads from each biological duplicate were merged and analyzed using NucleoATAC, with genomic regions set as ± 500 bp windows around Genrich peaks.

HINT-ATAC transcription factor footprinting

Quality-filtered reads from each biological duplicate were merged and analyzed using rgt-hint footprinting. We used the JASPAR2020 database and converted the position weight matrices from JASPAR format into a simple matrix format expected by RGT-HINT using the R package ‘universalmotif’ and generated a ‘.mtf’ file to store database information (see supplementary Materials and Methods). For enrichment analyses, we used bedtools random to generate 13 million random 20 bp sequences, as this was the average footprint size of genuine footprints detected by RGT-HINT in our data. This set of random sequences was used as background for our enrichment analyses. For cluster-specific enrichment analyses, we collated all unique transcription factor footprints from all developmental stages for each cluster (e.g. all footprints across S13, S14, etc. for all peaks in a given cluster) and compared enrichment levels with our randomly generated background.

Candidate reporter selection approaches

Careful examination of developmentally important genes revealed that many are surrounded by large numbers of peaks (>10) spread over large genomic distances. To further filter our Omni-ATAC peaks, we identified regions of sequence conservation to another amphipod crustacean, *Hyalella azteca* (Poynton et al., 2018), using the VISTA sequence alignment software (Ratnere and Dubchak, 2009) (Fig. S15A). *Hyalella* serves as a useful comparison to *Parhyale*, as its genome size is smaller (1.05 Gb) but, as it is also an amphipod crustacean, we expect that key developmental regulatory elements might have some level of sequence conservation.

Among our Omni-ATAC peaks, we took two strategies to identify candidate reporter elements. In the first approach, we identified all peaks within 5 kb of mRNA starts. In doing so, we were able to identify a handful of genomic regions in which we were able to locate both a putative promoter and a candidate proximal enhancer. We also identified several putative promoters, which we tested in isolation. In the second approach, we examined the genomic regions around important developmental genes of

interest. Many of the genes we examined showed very large numbers of strong peaks, and were therefore intractable to thorough analysis. For the purposes of this study, we focused on three regions: the region around the *Engrailed-1* gene, the region around the *Sp-69* gene and the region around the *Heat shock protein 70* complex, where two previous cis-regulatory elements (PhMS and HS2a) had been identified.

Minos transposon cloning

Minos transposon reporter plasmids were cloned using Gibson homology-mediated cloning approaches and the New England Biolabs Gibson Assembly or NEBuilder kits. As a base plasmid, we used the pMi(ne1) plasmid, which contains the Hsp70 minimal promoter, a DsRed protein sequence and an SV40 3' UTR sequence, as well as two Minos inverted repeats. For plasmids containing the Hsp70 minimal promoter, the insert was integrated between the EcoRV and BglII restriction sites. For plasmids containing an endogenous promoter, the insert replaced the sequence between the EcoRV and NcoI restriction sites, thereby removing the Hsp70 minimal promoter. Cloned plasmids were Sanger sequenced to confirm a full DsRed open reading frame (ORF) and inclusion of desired genomic sequences.

Minos transposase assay

Minos transposase mRNA was generated using the Thermo Fisher Scientific mMACHINE T7 or T7 ULTRA kit using NotI-digested pBlueSK-MimRNA (Addgene plasmid #102535). mRNA and concentrated DNA were mixed into a final concentration of 1 $\mu\text{g}/\mu\text{l}$ in a solution of 0.1% phenol red in nuclease-free water. One- and two-cell *Parhyale* embryos were injected with ~3–5 μl of injection mix using a borosilicate glass capillary needle pulled using a Sutter P-80 or P-85 instrument. Embryos were raised until hatching and examined once per day from 3 dpf until 10 dpf using a Zeiss LSM780 confocal microscope to screen for DsRed fluorescence.

Acknowledgements

We thank Jase Gehring and Jenna Haines for providing tips and reagents for performing and troubleshooting ATAC-seq, and Kasia Oktaba for suggesting the technique. Jenna Haines and Shaked Afik also provided helpful programming references for performing ATAC-seq analyses. We are grateful to Aaron Pomerantz for helping to troubleshoot and perform Nanopore sequencing, and for tips on RNA-seq analysis. This work would have been much more challenging without fantastic sequencing support from the QB3 Biosciences Functional Genomics Lab, particularly Shana McDevitt, Karen Lundy and Justin Choi. We thank Gideon Harijanja, Lauren Zane and Thienkim Ho for their experimental support on early stages of this project that were ultimately not included in the manuscript. We are also grateful to Dan Rokhsar for helpful comments on the data analysis and manuscript. Thank you to the members of the Patel Lab, Craig Miller's lab, Iswar Hariharan's lab and Hernan Garcia's lab for providing feedback and suggestions to the manuscript, particularly Craig Miller, Tyler Square and Brandon Schlomann.

Competing interests

The authors declare no competing or financial interests.

Author contributions

Conceptualization: D.A.S.; Methodology: D.A.S., J.V.B., H.S.B.; Software: D.A.S.; Validation: D.A.S.; Formal analysis: D.A.S.; Investigation: D.A.S.; Resources: D.A.S., J.V.B., H.S.B.; Data curation: D.A.S., J.V.B., H.S.B.; Writing - original draft: D.A.S.; Writing - review & editing: D.A.S., J.V.B., H.S.B., N.H.P.; Visualization: D.A.S.; Supervision: D.A.S., N.H.P.; Project administration: N.H.P.; Funding acquisition: N.H.P.

Funding

D.A.S. was supported by the National Science Foundation (NSF) Division of Graduate Education, GRFP 2016230010. This work was funded by an NSF Division of Integrative Organismal Systems grant (1257379) to N.H.P.

Data availability

Omni-ATAC-seq raw reads are available at SRA: Bioproject PRJNA765106. RNA-seq raw reads are available at SRA: Bioproject PRJNA765726. Omni-ATAC downstream analyses are available at GEO under accession number GSE197886. RNA-seq downstream analyses are available at GEO under accession number GSE202252. Downstream analysis files for Omni-ATAC-seq include BigWig files for visualizing Omni-ATAC read pileups; stage-specific and combined Genrich peaks;

NucleoATAC signal, occupancy, insert and other miscellaneous data; HINT-ATAC transcription factor footprinting predictions and enrichment statistics; JASPAR2020 converted database and .mtf file for HINT-ATAC analyses; and differential accessibility analysis tables. Downstream analysis files for RNA-seq include BigWig files for visualizing RNA-seq read pileups; Nanopore alignment sequences; individual transcriptome GFF and FASTA files; eggNOG annotations and OrthoFinder annotations of transcripts; and RNA-seq differential expression analysis tables. In addition, the GEO database contains the updated Mikado genome annotation and assignment of peaks based on their position relative to genes, along with classification data for peaks as concordant or discordant in expression and accessibility. Code for analysis and visualizations is available at GitHub (https://github.com/mezarque/Parhyale_Genome_Resources). Software version numbers are cataloged in Table S3.

Peer review history

The peer review history is available online at <https://journals.biologists.com/dev/article-lookup/doi/10.1242/dev.200793>.

References

- Borsari, B., Villegas-Mirón, P., Pérez-Lluch, S., Turpin, I., Laayouni, H., Segarra-Casas, A., Bertranpetit, J., Guigó, R. and Acosta, S. (2021). Enhancers with tissue-specific activity are enriched in intronic regions. *Genome Res.* **31**, 1325-1336. doi:10.1101/gr.270371.120
- Brogaard, K., Xi, L., Wang, J.-P. and Widom, J. (2012). A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496-501. doi:10.1038/nature11142
- Browne, W. E., Price, A. L., Gerberding, M. and Patel, N. H. (2005). Stages of embryonic development in the amphipod crustacean, *Parhyale hawaiiensis*. *Genesis* **42**, 124-149. doi:10.1002/gene.20145
- Bruce, H. S. and Patel, N. H. (2020). Knockout of crustacean leg patterning genes suggests that insect wings and body walls evolved from ancient leg segments. *Nat. Ecol. Evol.* **4**, 1703-1712. doi:10.1038/s41559-020-01349-0
- Buenrostro, J. D., Wu, B., Chang, H. Y. and Greenleaf, W. J. (2015). ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1-21.29.9. doi:10.1002/0471142727.mb2129s109
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421. doi:10.1186/1471-2105-10-421
- Cazet, J. F., Cho, A. and Juliano, C. E. (2021). Generic injuries are sufficient to induce ectopic Wnt organizers in Hydra. *eLife* **10**, e60562. doi:10.7554/eLife.60562
- Chechik, G. and Koller, D. (2009). Timing of gene expression responses to environmental changes. *J. Comput. Biol.* **16**, 279-290. doi:10.1089/cmb.2008.13TT
- Christian, S.-E., Schultheis, D., Schwirz, J., Ströhlein, N., Troelensberg, N., Majumdar, U., Dao, V., Grossmann, D., Richter, T., Tech, M. et al. (2015). The iBeetle large-scale RNAi screen reveals gene functions for insect development and physiology. *Nat. Commun.* **6**, 7822. doi:10.1038/ncomms8822
- Corces, M. R., Trevino, A. E., Hamilton, E. G., Greenside, P. G., Sinnott-Armstrong, N. A., Vesuna, S., Satpathy, A. T., Rubin, A. J., Montine, K. S., Wu, B. et al. (2017). An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959-962. doi:10.1038/nmeth.4396
- Crawford, K., Diaz Quiroz, J. F., Koenig, K. M., Ahuja, N., Albertin, C. B. and Rosenthal, J. J. C. (2020). Highly efficient knockout of a squid pigmentation gene. *Curr. Biol.* **30**, 3484-3490.e4. doi:10.1016/j.cub.2020.06.099
- Emms, D. M. and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238. doi:10.1186/s13059-019-1832-y
- Fischer, D. S., Theis, F. J. and Yosef, N. (2018). Impulse model-based differential expression analysis of time course sequencing data. *Nucleic Acids Res.* **46**, e119. doi:10.1093/nar/gky675
- Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Corread, S., Gheorghe, M., Baranašić, D. et al. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87-D92. doi:10.1093/nar/gkz1001
- Gatzmann, F., Falckenhayn, C., Gutekunst, J., Hanna, K., Raddatz, G., Carneiro, V. C. and Lyko, F. (2018). The methylome of the marbled crayfish links gene body methylation to stable expression of poorly accessible genes. *Epigenet. Chromatin* **11**, 57. doi:10.1186/s13072-018-0229-6
- Gehrke, A. R., Schneider, I., de la Calle-Mustienes, E., Tena, J. J., Gomez-Marín, C., Chandran, M., Nakamura, T., Braasch, I., Postlethwait, J. H., Gómez-Skarmeta, J. L. et al. (2015). Deep conservation of wrist and digit enhancers in fish. *Proc. Natl. Acad. Sci. USA* **112**, 803-808. doi:10.1073/pnas.1420208112
- Gehrke, A. R., Neverett, E., Luo, Y.-J., Brandt, A., Ricci, L., Hulet, R. E., Gompers, A., Ruby, J. G., Rokhsar, D. S., Reddien, P. W. et al. (2019). Acoel genome reveals the regulatory landscape of whole-body regeneration. *Science* **363**, eaau6173. doi:10.1126/science.aau6173
- Gilbert, D. G. (2019). Longest protein, longest transcript or most expression, for accurate gene reconstruction of transcriptomes? *bioRxiv*, 829184. doi:10.1101/829184
- Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. and Lieb, J. D. (2007). FAIRE (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877-885. doi:10.1101/gr.5533506
- Gisselbrecht, S. S., Palagi, A., Kurland, J. V., Rogers, J. M., Ozadam, H., Zhan, Y., Dekker, J. and Bulyk, M. L. (2020). Transcriptional silencers in drosophila serve a dual role as transcriptional enhancers in alternate cellular contexts. *Mol. Cell* **77**, 324-337.e8. doi:10.1016/j.molcel.2019.10.004
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q. et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-652. doi:10.1038/nbt.1883
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M. et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494-1512. doi:10.1038/nprot.2013.084
- Halfon, M. S. (2020). Silencers, enhancers, and the multifunctional regulatory genome. *Trends Genet.* **36**, 149-151. doi:10.1016/j.tig.2019.12.005
- Hinton, G. E. and Roweis, S. (2003). Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems* (ed. S. Becker, S. Thrun and K. Obermayer), pp. 857-864. MIT Press.
- Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattai, T., Jensen, L. J. et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309-D314. doi:10.1093/nar/gky1085
- Kao, D., Lai, A. G., Stamatakis, E., Rosic, S., Konstantinides, N., Jarvis, E., Donfrancesco, A., Natalia, P.-S., Semon, M., Grillo, M. et al. (2016). The genome of the crustacean *Parhyale hawaiiensis*, a model for animal development, regeneration, immunity and lignocellulose digestion. *Elife* **5**, 065789. doi:10.7554/elif.20062
- Kissane, S., Dhandapani, V. and Orsini, L. (2021). Protocol for assay of transposase accessible chromatin sequencing in non-model species. *STAR Protoc.* **2**, 100341. doi:10.1016/j.xpro.2021.100341
- Konstantinides, N. and Averof, M. (2014). A common cellular basis for muscle regeneration in arthropods and vertebrates. *Science* **343**, 788-791. doi:10.1126/science.1243529
- Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L. and Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278. doi:10.1186/s13059-019-1910-1
- Kumar, L. and E Futschik, M. (2007). Mfuzz: a software package for soft clustering of microarray data. *Bioinformatics* **2**, 5-7. doi:10.6026/97320630002005
- Lai, Y.-T., Deem, K. D., Borrás-Castells, F., Sambrani, N., Rudolf, H., Suryamohan, K., El-Sherif, E., Halfon, M. S., McKay, D. J. and Tomoyasu, Y. (2018). Enhancer identification and activity evaluation in the red flour beetle, *Tribolium castaneum*. *Development* **145**, dev160663. doi:10.1242/dev.160663
- Langmead, B. and Salzberg, S. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357-359. doi:10.1038/nmeth.1923
- Legendre, P. and Legendre, L. (2012). *Numerical Ecology*. Elsevier.
- Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100. doi:10.1093/bioinformatics/bty191
- Li, Y., Chen, C., Kaye, A. M. and Wasserman, W. W. (2015). The identification of cis-regulatory elements: a review from a machine learning perspective. *Biosystems* **138**, 6-17. doi:10.1016/j.biosystems.2015.10.002
- Li, Z., Schulz, M. H., Look, T., Begemann, M., Zenke, M. and Costa, I. G. (2019). Identification of transcription factor binding sites using ATAC-seq. *Genome Biol.* **20**, 45. doi:10.1186/s13059-019-1642-2
- Liubicich, D. (2007). The role of Hox genes in crustacean development and appendage specialization. PhD thesis, University of California, Berkeley, Berkeley, CA.
- Love, M. I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550. doi:10.1186/s13059-014-0550-8
- Ludwig, M., Patel, N. and Kreitman, M. (1998). Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development* **125**, 949-958. doi:10.1242/dev.125.5.94
- Mahony, S. and Benos, P. V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* **35**, W253-W258. doi:10.1093/nar/gkm272
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* **17**, 10-12. doi:10.14806/ej.17.1.200
- Martin, A., Serano, J. M., Jarvis, E., Bruce, H. S., Wang, J., Ray, S., Barker, C. A., O'Connell, L. C. and Patel, N. H. (2016). CRISPR/Cas9 mutagenesis reveals versatile roles of hox genes in crustacean limb specification and evolution. *Curr. Biol.* **26**, 14-26. doi:10.1016/j.cub.2015.11.021
- Mito, T., Nakamura, T., Bando, T., Ohuchi, H. and Noji, S. (2011). The advent of RNA interference in entomology. *Entomol. Sci.* **14**, 1-8. doi:10.1111/j.1479-8298.2010.00408.x

- Morton, J. T., Toran, L., Edlund, A., Metcalf, J. L., Lauber, C., Knight, R. and Jansson, J. K. (2017). Uncovering the horseshoe effect in microbial analyses. *mSystems* **2**, e00166-16. doi:10.1128/mSystems.00166-16
- Paris, M., Wolff, C., Patel, N. H. and Averof, M. (2022). Chapter Eight - The crustacean model *Parhyale hawaiiensis*. In *Current Topics in Developmental Biology* (ed. B. Goldstein and M. Srivastava), pp. 199-230. Academic Press.
- Pavlopoulos, A. and Averof, M. (2005). Establishing genetic transformation for comparative developmental studies in the crustacean *Parhyale hawaiiensis*. *Proc. Natl. Acad. Sci. USA* **102**, 7888-7893. doi:10.1073/pnas.0501101102
- Pavlopoulos, A., Kontarakis, Z., Liubicich, D. M., Serano, J. M., Akam, M., Patel, N. H. and Averof, M. (2009). Probing the evolution of appendage specialization by Hox gene misexpression in an emerging model crustacean. *Proc. Natl. Acad. Sci. USA* **106**, 13897-13902. doi:10.1073/pnas.0902804106
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825-2830.
- Pérez-Zamorano, B., Rosas-Madrugal, S., Lozano, O. A. M., Castillo Méndez, M. and Valverde-Garduño, V. (2017). Identification of cis-regulatory sequences reveals potential participation of *lola* and *Deaf1* transcription factors in *Anopheles gambiae* innate immune response. *PLoS One* **12**, e0186435. doi:10.1371/journal.pone.0186435
- Picelli, S., Björklund, Å. K., Reinius, B., Sagasser, S., Winberg, G. and Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively-scaled sequencing projects. *Genome Res.* **24**, 2033-2040. doi:10.1101/gr.177881.114
- Podani, J. and Miklós, I. (2002). Resemblance coefficients and the horseshoe effect in principal coordinates analysis. *Ecology* **83**, 3331-3343. doi:10.1890/0012-9658(2002)083[3331:RCATHE]2.0.CO;2
- Poynton, H. C., Hasenbein, S., Benoit, J. B., Sepulveda, M. S., Poelchau, M. F., Hughes, D. S. T., Murali, S. C., Chen, S., Glastad, K. M., Goodisman, M. A. D. et al. (2018). The toxicogenome of *hyalella azteca*: a model for sediment ecotoxicology and evolutionary toxicology. *Environ. Sci. Technol.* **52**, 6009-6022. doi:10.1021/acs.est.8b00837
- Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842. doi:10.1093/bioinformatics/btq033
- Radman-Livaja, M. and Rando, O. J. (2010). Nucleosome positioning: how is it established, and why does it matter? *Dev. Biol.* **339**, 258-266. doi:10.1016/j.ydbio.2009.06.012
- Ramos, A. P., Gustafsson, O., Labert, N., Salecker, I., Nilsson, D.-E. and Averof, M. (2019). Analysis of the genetically tractable crustacean *Parhyale hawaiiensis* reveals the organisation of a sensory system for low-resolution vision. *BMC Biol.* **17**, 67. doi:10.1186/s12915-019-0676-y
- Rasys, A. M., Park, S., Ball, R. E., Alcalá, A. J., Lauderdale, J. D. and Menke, D. B. (2019). CRISPR-Cas9 gene editing in lizards through microinjection of unfertilized oocytes. *Cell Rep.* **28**, 2288-2292.e3. doi:10.1016/j.celrep.2019.07.089
- Ratnere, I. and Dubchak, I. (2009). Obtaining comparative genomic data with the VISTA family of computational tools. *Curr. Protoc. Bioinform.* **26**, 10.6.1-10.6.17. doi:10.1002/0471250953.bi1006s26
- Rehm, E. J., Hannibal, R. L., Chaw, R. C., Vargas-Vila, M. A. and Patel, N. H. (2009). Fixation and dissection of *parhyale hawaiiensis* embryos. *Cold Spring Harb. Protoc.* **2009**, pdb.prot5127. doi:10.1101/pdb.prot5127
- Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140. doi:10.1093/bioinformatics/btp616
- Rouhana, L., Weiss, J. A., Forsthoefel, D. J., Lee, H., King, R. S., Inoue, T., Shibata, N., Agata, K. and Newmark, P. A. (2013). RNA interference by feeding in vitro-synthesized double-stranded RNA to planarians: methodology and dynamics. *Dev. Dyn.* **242**, 718-730. doi:10.1002/dvdy.23950
- Schep, A. N., Buenrostro, J. D., Denny, S. K., Schwartz, K., Sherlock, G. and Greenleaf, W. J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* **25**, 1757-1770. doi:10.1101/gr.192294.115
- Sharma, P. P., Schwager, E. E., Giribet, G., Jockusch, E. L. and Extavour, C. G. (2014). Whole-body acoel regeneration is controlled by Wnt and Bmp-admp signaling. *Curr. Biol.* **24**, 1107-1113. doi:10.1016/j.cub.2014.03.042
- Stamatakis, E. and Pavlopoulos, A. (2016). Non-insect crustacean models in developmental genetics including an encyrtid to *Parhyale hawaiiensis*. *Curr. Opin. Genet. Dev.* **39**, 149-156. doi:10.1016/j.gde.2016.07.004
- Sun, D. A. and Patel, N. H. (2019). The amphipod crustacean *Parhyale hawaiiensis*: An emerging comparative model of arthropod development, evolution, and regeneration. *Wiley Interdiscip. Rev. Dev. Biol.* **8**, e355. doi:10.1002/wdev.355
- The UniProt Consortium. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506-D515. doi:10.1093/nar/gky1049
- Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579-2605. doi:10.1161/1.3267037
- Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. and Swarbreck, D. (2018). Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *GigaScience* **7**, giy093. doi:10.1093/gigascience/giy093
- Wang, Q., Gu, L., Adey, A., Radlwimmer, B., Wang, W., Hovestadt, V., Bähr, M., Wolf, S., Shendure, J., Eils, R. et al. (2013). Tagmentation-based whole-genome bisulfite sequencing. *Nat. Protoc.* **8**, 2022-2032. doi:10.1038/nprot.2013.118
- Wasserman, W. W. and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276-287. doi:10.1038/nrg1315
- Yan, F., Powell, D. R., Curtis, D. J. and Wong, N. C. (2020). From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol.* **21**, 22. doi:10.1186/s13059-020-1929-3
- Yosef, N. and Regev, A. (2011). Impulse control: temporal dynamics in gene transcription. *Cell* **144**, 886-896. doi:10.1016/j.cell.2011.02.015